V. DYACHENKO

NOTIONS FONDAMENTALES DU CALCUL NUMÉRIQUE

V. Diatchenko

NOTIONS FONDAMENTALES DU CALCUL NUMÉRIQUE

Editions Mir • Moscou

PRÉFACE

Les possibilités offertes par les ordinateurs pour la réalisation des expériences de calcul ont accéléré le processus de mathématisation scientifique et technique. Un nombre de plus en plus grand de professions exigent des connaissances mathématiques approfondies. Pour subvenir aux besoins de l'enseignement, de nombreux mémentos et monographies ont été édités. Le présent ouvrage appartient à cette série.

Cet ouvrage est consacré aux questions liées à l'élaboration des méthodes de solution numérique des problèmes de mécanique et de physique mathématique. Actuellement, la théorie est en pleine essor et est loin d'être achevée. Néanmoins un certain nombre d'idées et de notions assez simples sont devenues fondamentales, elles sont exposées dans le présent ouvrage. Il est évident que la connaissance de l'alphabet des méthodes numériques ne saurait faire du lecteur un calculateur qualifié. Il faut à cet effet une étude plus profonde de la littérature et surtout une grande expérience personnelle de résolution des problèmes concrets qui deviennent bien plus efficaces lorsque les fondements de la théorie sont clairs.

La diversité des professions et du niveau de préparation du lecteur ont obligé l'auteur à renoncer à un exposé formel, traditionnel pour la littérature mathématique. Il est ainsi possible que certains mathématiciens professionnels raffinés trouvent cet ouvrage banal et peut-être même vulgaire, et nous sommes d'accord avec eux. Mais un « armement » mathématique complet, tenant compte de toutes les éventualités logiques n'est indispensable que si l'on ne dispose que de la logique mathématique. Mais il n'en est pas ainsi pour la majorité des objectifs réels, ce qui permet d'écrire des ouvrages tels que celui-ci. D'un autre coté nous désirons que ce livre soit lu, nous avons donc essayé de faire de telle

sorte que le lecteur n'ait pas de raisons pour le rejeter.

L'exposé est mené au niveau de rigueur « physique ». L'auteur tient compte moins de la préparation mathématique du lecteur que de son bon sens et de son intelligence. En général, l'étude de telle ou telle question est menée sur un exemple typique simple, puis la possibilité de généralisation des résultats obtenus à des cas plus compliqués est envisagée.

A la fin de chaque paragraphe on trouve des problèmes qui sont de difficultés différentes. La solution de la majorité de ces problèmes exige une compréhension claire des principes exposés dans le texte et la faculté de les développer indépendamment.

Les trois premiers paragraphes servent d'introduction et concernent les questions classiques de calcul numérique, à savoir les méthodes itératives de résolution des équations, les formules d'interpolation et de quadrature, l'intégration numérique des équations différentielles ordinaires.

Les autres paragraphes formant le contenu essentiel de l'ouvrage sont consacrés aux méthodes numériques de résolution des équations aux dérivées partielles, à l'élaboration et à l'étude des algorithmes de calcul correspondants. On y envisage les procédés concrets d'approximation et de stabilité des problèmes aux différences, les propriétés des schémas explicites et implicites, les méthodes d'élaboration des formules de calcul, les méthodes de résolution des systèmes d'équations aux différences, la possibilité de réalisation des algorithmes, etc.

Nous proposons donc au lecteur une théorie simplifiée des méthodes numériques modernes de résolution des problèmes de mécanique, de dynamique des gaz et de physique mathématique donnant lieu à l'intégration des équations différentielles.

Cet ouvrage est le fruit de longues années de travail de l'auteur dans le domaine des mathématiques appliquées et de

conférences faites à l'Ecole Physico-technique de Moscou. Il peut être utile aux étudiants des facultés des sciences et des Grandes Ecoles pour une première connaissance des méthodes de calcul numérique.

Les questions abordées dans les trois premiers paragraphes sont exposées dans de nombreux ouvrages et on peut les trouver en détail dans tous les cours de calcul approché.

Pour une étude plus approfondie des méthodes de résolution des équations aux dérivées partielles on peut recommander les ouvrages suivants:

1. Годунов С. К., Рябенький В. С. Введение в теорию разностных схем (Introduction à la théorie des schémas aux différences). Физматгиз, 1962.

2. Самарский А. А. Введение в теорию разностных схем (Introduction à la théorie des schémas aux différences). « Наука », 1971.

3. Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики (Méthode des pas fractionnaires de résolution des problèmes multidimensionnels de physique mathématique). « Наука », 1967.

4. Richtmyer R. Difference Methods for Initial-Value Problems. New York, 1957.

5. Wasow W., Forsythe G. Finite-difference Methods for Partial Differential Equations, New York, 1960.

L'auteur tend à remercier V. S. Riabenki pour son concours lors de la parution du présent ouvrage.

V. Diatchenko

On appelle méthode numérique de résolution d'un problème une certaine suite d'opérations sur des nombres, c'est-à-dire un algorithme de calcul dont le langage est formé par des chiffres et des opérations arithmétiques. Ce langage primitif permet de réaliser les méthodes numériques sur ordinateur, ce qui fait qu'elles sont un instrument puissant et universel de recherche.

Cependant les problèmes à résoudre se formulent en général dans le langage mathématique ordinaire (équations, fonctions, opérateurs différentiels, etc.). Ainsi l'élaboration d'une méthode numérique suppose obligatoirement la substitution, l'approximation du problème initial par un autre voisin formulé en termes de nombres et d'opérations arithmétiques. Malgré la diversité des méthodes de cette substitution certaines propriétés générales leur sont communes.

Considérons un exemple très simple. Il y a lieu de trouver la solution de l'équation suivante

$$x^2 - a = 0, \qquad a > 0, \tag{1}$$

c'est-à-dire d'extraire la racine carrée d'un nombre donné a. On peut évidemment écrire $x = \sqrt{a}$, mais le symbole $\sqrt{\ }$ ne résout pas le problème, ne donne pas une méthode de calcul de la grandeur x.

Nous allons procéder de la manière suivante. Nous allons nous donner une approximation initiale x_0 (par exemple $x_0 = 1$) et successivement à l'aide de la formule

$$x_n = \frac{1}{2}\left(x_{n-1} + \frac{a}{x_{n-1}}\right) \tag{2}$$

calculer les valeurs $x_1, x_2, \ldots$ Nous allons arrêter ce processus sur un certain $n = N$, le résultat obtenu x_N sera déclaré

être la solution approchée du problème initial (1), c'est-à-dire :

$$\sqrt{a} \sim x_N.$$

La possibilité de cette hypothèse dépend évidemment des exigences imposées à la précision de la solution, de la grandeur a et du paramètre N. Pour des conditions quelconques, il y a lieu de démontrer que, pour tout a, par un choix convenable de N on peut faire x_N aussi voisin que l'on veut de la valeur exacte $\sqrt{a}$.

Nous allons démontrer que notre alogrithme (2) satisfait à cette condition. Posons

$$\frac{x_n}{\sqrt{a}} = 1 + \varepsilon_n. \tag{3}$$

Divisions l'égalité (2) par $\sqrt{a}$ et substituons (3), il vient

$$1 + \varepsilon_n = \frac{1}{2}\left(1 + \varepsilon_{n-1} + \frac{1}{1+\varepsilon_{n-1}}\right),$$

d'où

$$\varepsilon_n = \frac{1}{2}\left(\varepsilon_{n-1} - 1 + \frac{1}{\varepsilon_{n-1}+1}\right) = \frac{1}{2}\,\frac{\varepsilon_{n-1}^2}{\varepsilon_{n-1}+1}. \tag{4}$$

Comme $1 + \varepsilon_0 = 1/\sqrt{a} > 0$, en vertu de la dernière égalité, tous les ε_n à partir du premier sont positifs. Ceci signifie que

$$\frac{\varepsilon_{n-1}}{\varepsilon_{n-1}+1} < 1.$$

Compte tenu de ce qui vient d'être dit, on obtient à partir de (4)

$$\varepsilon_n < \frac{1}{2}\,\varepsilon_{n-1}, \tag{5}$$

c'est-à-dire que ε_n décroît lorsque n augmente, et ceci plus rapidement qu'une progression géométrique de raison 1/2. Par conséquent

$$x_N \to \sqrt{a} \qquad \text{pour } N \to \infty, \tag{6}$$

notre assertion se trouve ainsi démontrée.

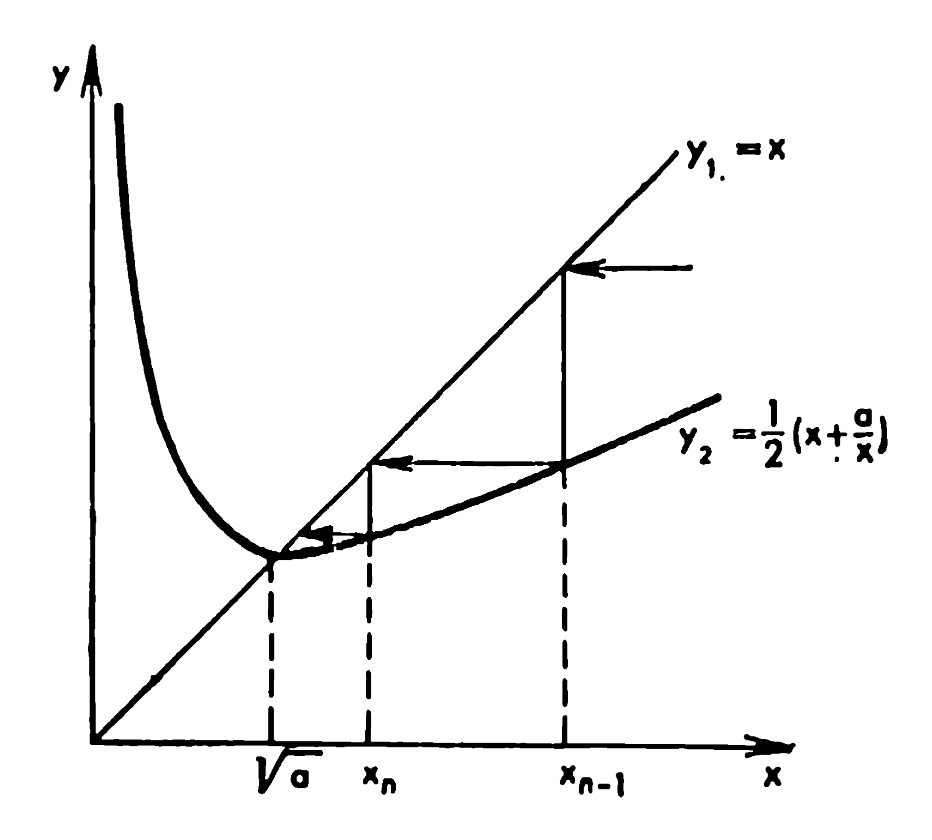


Fig. 1.

Sur la fig. 1 on peut voir le processus itératif (2). Les graphiques correspondant aux membres premier $y_1(x)$ et second $y_2(x)$ de (2). Comme de toute évidence $y_1(\sqrt{a}) = y_2(\sqrt{a})$, ces graphiques se rencontrent au point $x = \sqrt{a}$. Les itérations correspondant à la formule (2) équivalent à un mouvement le long de la ligne brisée représentée sur la figure entre $y_1(x)$ et $y_2(x)$. On peut ainsi se rendre compte une fois de plus de ce que les itérations convergent vers $\sqrt{a}$, pour $N \to \infty$.

Lors de l'étude de la convergence nous avons en une certaine mesure idéalisé l'algorithme en supposant que le calcul exact suivant la formule (2) est possible, mais ni l'homme, ni l'ordinateur ne peuvent opérer avec des nombres réels quelconques, les calculs s'effectuent toujours avec un nombre restreint de décimales et la précision du résultat ne saurait être supérieure à la précision du calcul. Il importe donc d'établir la relation existant entre ces précisions, de

voir si les erreurs apparaissant lors de l'arrondissement ne s'accumulent pas, ce qui aurait pour effet de donner un résultat dépourvu de toute valeur.

Bien qu'il soit presque évident que, dans notre exemple, ceci n'aura pas lieu, nous allons procéder à la vérification formelle. L'arrondissement revient en fait à remplacer la formule (2) par la formule

$$\tilde{x}_n = \frac{1}{2}\left(\tilde{x}_{n-1} + \frac{a}{\tilde{x}_{n-1}}\right)(1 + \delta_n), \tag{7}$$

où le facteur $1 + \delta_n$ tient compte de l'erreur d'arrondissement introduite lors du n-ième pas du calcul, $\tilde{x}_n$ étant la suite obtenue en fait. La grandeur $\delta \ll 1$ caractérise la précision des calculs. En remplaçant $\tilde{x}_n$ par $\sqrt{a}\,(1 + \varepsilon_n)$ on obtient au lieu de (4)

$$\varepsilon_n = \frac{1}{2}\,\frac{\varepsilon_{n-1}^2}{\varepsilon_{n-1} + 1}(1 + \delta_n) + \delta_n .$$

On peut voir qu'avec l'augmentation de n ε_n décroît jusqu'à une grandeur de l'ordre de δ_n, c'est-à-dire que la précision du résultat correspond à la précision des calculs.

Malgré sa simplicité, l'exemple envisagé met en évidence d'une manière très claire les principes généraux suivants caractérisant toutes les méthodes de calcul.

Primo, le problème initial (1) est remplacé par un autre problème, à savoir un algorithme de calcul (2).

Secondo, le problème (2) contient un paramètre N ne figurant pas dans le problème initial.

Tertio, par un choix convenable de ce paramètre on peut en principe rendre la solution x_N du second problème aussi voisine que l'on veut de la solution du premier $\sqrt{a}$.

Enfin quarto, si l'algorithme est réalisé avec une précision insuffisante, vu l'arrondissement, ceci n'a en fait pas d'influence sur ses propriétés.

Chapitre premier

§ 1. CALCUL DES RACINES D'UNE ÉQUATION

Soit à trouver la racine réelle de l'équation

$$f(x) = 0. \tag{8}$$

Nous allons supposer que cette racine se trouve à l'intérieur d'un certain intervalle donné (celui-ci peut être grand).

Le problème (1) étudié ci-dessus est un cas particulier de celui envisagé ici. La méthode d'itérations utilisée peut être étendue au cas général de l'équation (8). A cet effet il y a lieu de réaliser un processus itératif du type

$$x_n = \varphi(x_{n-1}) \tag{9}$$

(comparer avec (2)) convergeant vers la racine de l'équation (8) que nous désignerons par X.

Nous allons voir les conditions auxquelles doit satisfaire la fonction $\varphi(x)$ et certaines méthodes permettant de la construire.

Supposons que le processus d'itérations (9) converge, c'est-à-dire que $x_n \to X$ pour $n \to \infty$. En passant à la limite dans les premier et second membres de (9), on obtient $X = \varphi(X)$, soit $x = X$ doit être une racine commune à l'équation (8) et à

$$x = \varphi(x). \tag{10}$$

Donc pour obtenir la formule d'itérations, il suffit simplement d'écrire l'équation (8) sous la forme (10) (comparer (1) et (2)), ce que l'on peut de toute évidence faire de différentes

manières. Ainsi l'équation (1) peut, en plus de la forme (2), s'écrire comme suit :

$$x = \frac{a}{x}$$

ou

$$x = 2x - \frac{a}{x}.$$

Il est facile de voir que ni l'une ni l'autre de ces formules ne peut être utilisée pour des itérations. La première donne $x_1 = a/x_0$, $x_2 = x_0$, $x_3 = a/x_0$, . . ., et la seconde donne naissance à une suite x_n tendant vers l'infini. Ainsi, dans la plupart des cas, en écrivant (8) sous la forme (10), on n'obtient pas le résultat désiré.

Nous allons voir quelles sont les propriétés de $\varphi(x)$ qui influent sur la convergence du processus. Comme la convergence signifie que $x_n - X \to 0$ pour $n \to \infty$, il faut que $|x_n - X|$ diminue lorsque n augmente. Supposons que pour n quelconque on ait l'inégalité

$$|x_n - X| \leqslant \Theta |x_{n-1} - X|, \qquad \Theta < 1. \qquad (11)$$

Il est alors évident que $|x_n - X|$ décroît comme une progression géométrique de raison Θ, c'est-à-dire qu'il y a convergence. Substituons dans le premier membre de (11), au lieu de x_n et X, respectivement $\varphi(x_{n-1})$ et $\varphi(X)$. On obtient

$$|\varphi(x_{n-1}) - \varphi(X)| \leqslant \Theta |x_{n-1} - X|, \qquad \Theta < 1. \qquad (12)$$

Comme on ne connaît pas la racine X, on ne peut vérifier directement la dernière condition et il faut un peu la renforcer. Ci-dessus nous avons supposé que la racine était localisée à l'intérieur d'un certain intervalle. Si pour deux points quelconques x', x'' de cet intervalle la condition suivante

$$|\varphi(x'') - \varphi(x')| \leqslant \Theta |x'' - x'|, \qquad \Theta < 1, \qquad (13)$$

se trouve remplie, la condition (12) se trouve vérifiée à fortiori. L'application $x = \varphi(x)$ satisfaisant à (13) est dite

contractante (elle contracte le segment $x'' - x'$). Il est évident que la condition (13) est une condition suffisante pour la convergence du processus itératif (9).

Pour l'estimation de la grandeur Θ déterminant la vitesse de convergence, le plus simple est d'utiliser la formule évidente

$$\Theta = \max_x |\varphi'(x)|, \tag{14}$$

où max est pris sur l'intervalle de la localisation de la racine. Si l'application $x = \varphi(x)$ est contractante, cet intervalle diminuera lorsque n augmente, et on peut préciser alors la vitesse de convergence.

Ainsi, lorsque l'on écrit (8) sous la forme (10) tout en satisfaisant à la condition (13), on obtient un processus itératif convergeant vers la racine.

Il est évident que la vitesse de convergence détermine la qualité de tel ou tel choix de $\varphi(x)$. En ce sens il va de soi que le meilleur est celui pour lequel la grandeur Θ soit minimale.

Soit $\varphi(x)$ tel que les équations $f(x) = 0$ et $\varphi(x) - x = 0$ aient une racine commune X. Si cette racine n'est pas multiple, $f'(X) \neq 0$, au voisinage de cette racine, où la fonction $f(x)$ n'a pas d'autres racines, le rapport

$$\frac{\varphi(x) - x}{f(x)} = r(x)$$

est alors une fonction bornée. A chaque fonction $\varphi(x)$ correspond $r(x)$ et inversement chaque $r(x)$ bornée donne naissance à

$$\varphi(x) = x + r(x)f(x). \tag{15}$$

Ce qui nous intéresse, c'est $\varphi(x)$ pour laquelle $|\varphi'(x)|$ est minimale (en vertu de (14)). En dérivant l'expression (15), on a

$$\varphi' = 1 + r(x)f'(x) + r'(x)f(x).$$

Au voisinage de la racine, la grandeur $f(x)$ est minimale, nous pouvons ainsi négliger le dernier terme et prendre l'expression restante égale à zéro. Ceci donne

$$r(x) = -\frac{1}{f'(x)},$$

c'est-à-dire, en vertu de (15), la fonction

$$\varphi(x) = x - \frac{f(x)}{f'(x)}, \tag{16}$$

donnant naissance au processus d'itérations, appelé *méthode de Newton*

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}. \tag{17}$$

La convergence de ce processus est très rapide car

$$\varphi'(x) = \frac{f''(x)}{(f'(x))^2} f(x) \to 0$$

au fur et à mesure que l'on s'approche de la racine.

On peut obtenir la formule (17) également d'une autre manière (fig. 2), à savoir, à partir d'une certaine approximation pour la racine x_{n-1}, lors du calcul de l'approximation suivante x_n nous allons, au lieu de l'équation $f(x) = 0$, considérer l'équation linéaire

$$f(x_{n-1}) + f'(x_{n-1})(x - x_{n-1}) = 0,$$

ne tenant compte que des deux premiers termes du développement de la fonction $f(x)$ en série de Taylor au voisinage du point x_{n-1}. En résolvant cette équation par rapport à x, on obtient la formule (17).

Les avantages de la méthode de Newton apparaissent seulement au voisinage de la racine, où φ' est petit. C'est pourquoi dans les calculs il y a lieu, avant tout, de localiser la racine par une méthode plus simple et plus grossière, puis pour obtenir rapidement une grande précision, utiliser

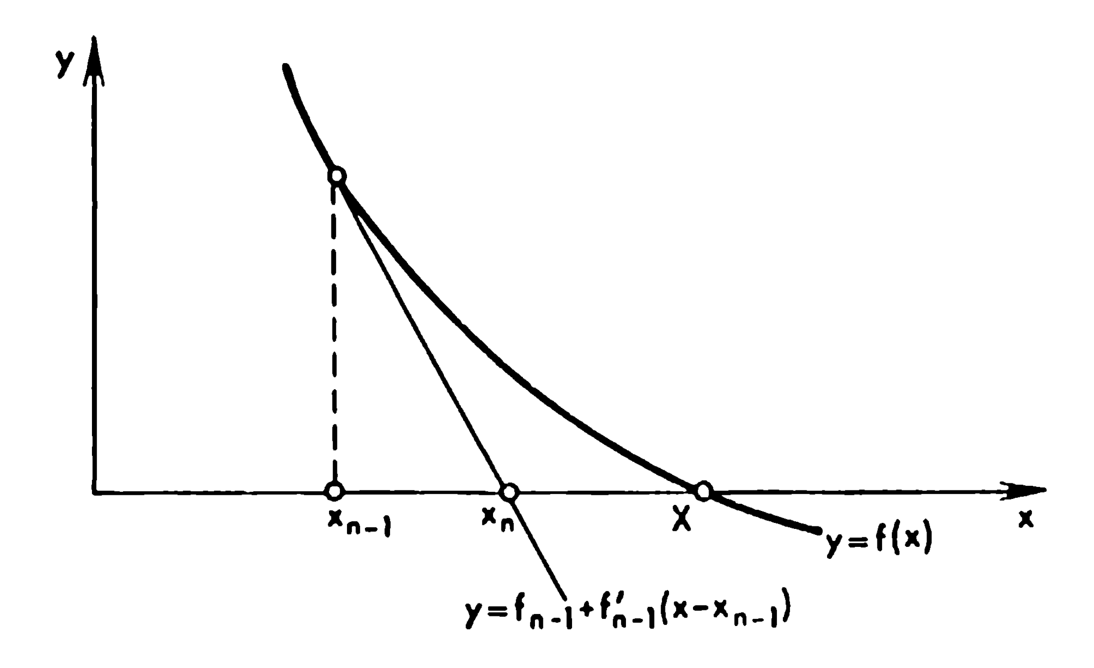


Fig. 2

la méthode de Newton. La précision des calculs (nombre de décimales retenues) doit être variable, il y a lieu de l'augmenter au fur et à mesure que l'on se rapproche de la racine. Comme la transition de x_{n-1} à x_n est simultanément une vérification de l'équation $x = \varphi(x)$, équivalente à $f(x) = 0$, la précision atteinte peut être estimée d'après le nombre de décimales restant inchangées lors de cette transition.

La méthode itérative de recherche des racines exposée, utilisant des applications contractantes, peut sans modifications de principe être généralisée à de nombreux problèmes plus compliqués.

Soit à résoudre un système de M équations à M inconnues

$$f^{(m)}(x^{(1)}, x^{(2)}, \ldots, x^{(M)}) = 0, \qquad m = 1, 2, \ldots M. \quad (18)$$

Désignons par f le vecteur de composantes $f^{(1)}, f^{(2)}, \ldots$ $\ldots, f^{(M)}$ et par x l'ensemble $x^{(1)}, x^{(2)}, \ldots, x^{(M)}$ et écrivons le système (18) sous la forme d'une équation vectorielle, soit

$$f(x) = 0 \quad (19)$$

dont la forme coïncide avec (8). Tout ce qui vient d'être dit au sujet de l'équation (8), peut être transposé au systè-

me (19), il faut seulement, compte tenu du caractère vectoriel des grandeurs, attribuer un sens nouveau aux désignations utilisées.

Pour l'estimation de la grandeur d'un scalaire on utilise son module, pour un vecteur on utilise sa *norme*. Nous définirons cette dernière comme le maximum du module des composantes

$$\| x \| = \max_m | x^{(m)} |. \tag{20}$$

Il est évident que si la norme d'un vecteur est égale à zéro ceci signifie que toutes ses composantes sont nulles.

$$x = \varphi (x), \tag{21}$$

où φ est le vecteur de composantes $\varphi^{(1)}, \varphi^{(2)}, \ldots, \varphi^{(m)}$. Tout comme précédemment la formule (21) donne naissance à un processus itératif qui converge si l'application (21) est contractante, c'est-à-dire

$$\| \varphi (x'') - \varphi (x') \| \leqslant \Theta \| x'' - x' \|, \qquad \Theta < 1. \tag{22}$$

Nous allons procéder à l'estimation de la grandeur Θ. A cet effet considérons la fonction $\varphi^{(m)} (x)$ sur la droite joignant x' et x''. On obtient une fonction de l'argument scalaire s

$$\psi (s) = \varphi^{(m)} (x' + s (x'' - x')). \tag{23}$$

Les points x' et x'' correspondent aux valeurs $s = 0$ et $s = 1$. Il est évident que

$$\varphi^{(m)} (x'') - \varphi^{(m)} (x') = \psi (1) - \psi (0) = \psi' (\tilde{s}), \tag{24}$$

avec $0 \leqslant \tilde{s} \leqslant 1$. En dérivant (23) par rapport à s, on obtient

$$\psi' (s) = \sum_{k=1}^{M} \frac{\partial \varphi^{(m)}}{\partial x^{(k)}} ((x^{(k)})'' - (x^{(k)})')$$

et compte tenu de (24)

$$|\varphi^{(m)}(x'')-\varphi^{(m)}(x')|\leqslant$$
$$\leqslant\max_x\sum_{k=1}^{M}\left|\frac{\partial\varphi^{(m)}}{\partial x^{(k)}}\right|\max_k|(x^{(k)})''-(x^{(k)})'|.$$

Compte tenu de la définition de la norme (20), on obtient

$$\|\varphi(x'')-\varphi(x')\|\leqslant\max_m\max_x\sum_{k=1}^{M}\left|\frac{\partial\varphi^{(m)}}{\partial x^{(k)}}\right|\|x''-x'\|. \quad (25)$$

Il est évident que maintenant c'est la matrice des dérivées $\frac{\partial\varphi^{(m)}}{\partial x^{(k)}}$ du vecteur fonction par rapport au vecteur argument qui joue le rôle de φ'. Si l'on définit maintenant la norme de cette matrice par la formule

$$\|\varphi'\|=\max_m\sum_k\left|\frac{\partial\varphi^{(m)}}{\partial x^{(k)}}\right|, \quad (26)$$

pour l'estimation de la grandeur Θ dans l'expression (22), en vertu de (25), on peut utiliser l'égalité

$$\Theta=\max_x\|\varphi'(x)\|, \quad (27)$$

qui est une généralisation de (14) au cas vectoriel.

Puis on peut, comme précédemment, construire $\varphi(x)$ à l'aide de (15)

$$\varphi(x)=x+r(x)f(x), \quad (28)$$

où $r(x)$ est une matrice quelconque de fonctions. Pour

$$r(x)=-(f'(x))^{-1}, \quad (29)$$

où $(f')^{-1}$ est la matrice inverse à la matrice des dérivées $f'=\frac{\partial f^{(m)}}{\partial x^{(k)}}$, on obtient la méthode de Newton. Cette méthode conserve également ici ses avantages car la matrice qui lui

correspond φ' sera nulle au point de la racine. Il est facile de s'en rendre compte en dérivant (28) et en tenant compte de (29). Maintenant, en utilisant la méthode de Newton, au lieu de diviser par f' il suffit d'inverser la matrice f', c'est-à-dire lors de chaque itération de résoudre le système d'équations linéaires suivant

$$f(x_{n-1}) + f'(x_{n-1})(x - x_{n-1}) = 0.$$

Lorsque l'ordre du système M est élevé, ceci peut être assez compliqué. C'est pourquoi généralement on n'utilise cette méthode que dans le cas où l'on dispose d'une bonne approximation pour les racines, lorsque une ou deux itérations donnent une augmentation notable de la précision.

Il y a lieu de faire la remarque suivante concernant les systèmes d'équations linéaires. Si on a besoin de résoudre un système du type général il n'y a rien de mieux que la méthode bien connue d'exclusion. Cependant en utilisant les propriétés spécifiques de certains systèmes particuliers on arrive souvent à des méthodes itératives efficaces.

Problèmes

1. Construire le processus d'itérations permettant de résoudre la racine d'ordre p.
2. Trouver la rapidité de la convergence de la méthode de Newton au voisinage de la racine multiple d'ordre k.
3. Construire le processus d'itérations de la résolution de l'équation $ax + b = 0$ où $0{,}1 < a < 1$, sans utiliser l'opération de division.
4. Soit le système d'équations linéaires $Ax + b = 0$ de matrice A dont toutes les valeurs propres sont réelles, différentes, positives et se trouvent avec certitude à l'intérieur d'un certain intervalle donné (α, β). Choisir le nombre r de telle sorte que le processus itératif

$$x_n = x_{n-1} + r(Ax_{n-1} + b)$$

converge le plus rapidement possible.

§ 2. FONCTIONS ET TABLES

L'une des notions mathématiques fondamentales est celle de fonction, celle-ci est utilisée donc dans la formulation de la majorité des problèmes. Dans le langage des méthodes numériques une fonction ne peut être, de toute évidence, représentée que par des suites numériques. Différentes méthodes de représentation sont possibles, par exemple, à l'aide des coefficients d'une série par rapport à des fonctions ou des valeurs des paramètres dans une formule d'un type donné. La représentation la plus courante et la plus universelle d'une fonction est une table de valeurs. Ainsi, toutes les fonctions élémentaires sin x, ln x, etc., sont des tables.

Considérons la fonction $F(x)$ et la table correspondante, c'est-à-dire deux suites x_k, f_k ($k = 0, 1, 2, \ldots$). Nous allons estimer la conformité de la fonction et de la table par la précision avec laquelle la table x_k, f_k permet de restituer la valeur de $F(x)$ en un point quelconque x.

Il est clair que cette précision dépend de la proximité de f_k à $F(x_k)$ et de la densité des nœuds de la table des x_k. Le rôle du premier facteur est évident : l'écart de f_k à $F(x_k)$ donne la limite de la précision de réproduction qu'il est impossible de surpasser sans changer la table. En idéalisant le problème nous allons poser

$$f_k = F(x_k), \quad k = 0, 1, 2, \ldots \tag{30}$$

Dans ce cas plus la table est détaillée, mieux elle décrit les particularités du comportement de la fonction $F(x)$, plus grande est la précision de restitution, celle-ci dépendant évidemment de la méthode de restitution.

Soit, par exemple, la méthode la plus grossière consistant en ce que chaque valeur de f_k est attribuée à tout l'intervalle adjacent $x_k \leqslant x < x_{k+1}$ (fig. 3), c'est-à-dire que pour calculer $F(x)$ on utilise une fonction constante par morceaux

$$P_0(x) = f_k, \quad x_k \leqslant x < x_{k+1}, \quad k = 0, 1, 2, \ldots \tag{31}$$

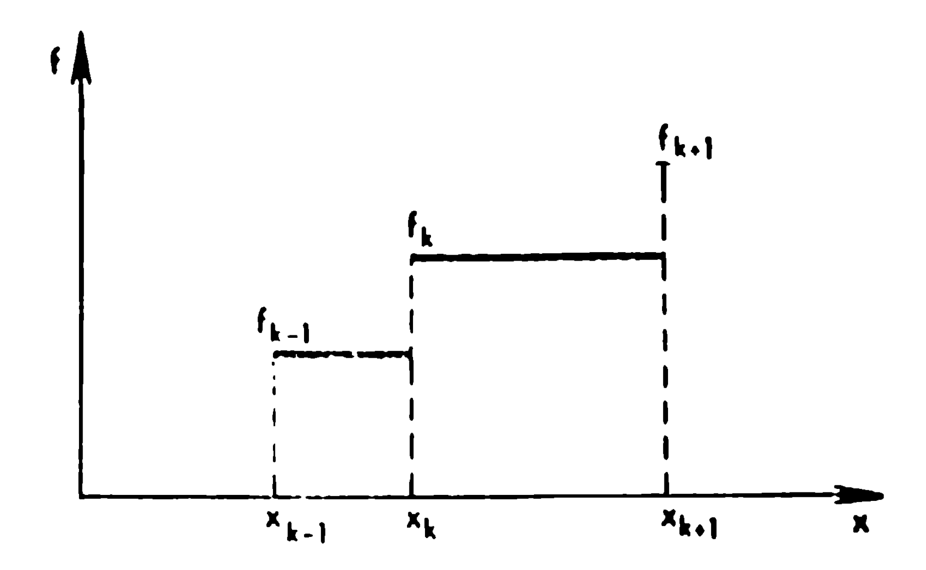


Fig. 3

Pour estimer la valeur de l'erreur $P_0(x) - F(x)$ il faut avoir en plus de (30) une information supplémentaire sur la fonction $F(x)$. Supposons que $F(x)$ soit une fonction lisse à dérivée continue. Sur l'intervalle (x_k, x_{k+1}) on peut alors la représenter comme suit:

$$F(x) = F(x_k) + F'(\xi(x))(x - x_k).$$

On voit immédiatement que sur cet intervalle on a

$$|P_0(x) - F(x)| \leqslant \max |F'|(x_{k+1} - x_k), \qquad (32)$$

c'est-à-dire que l'erreur est *de l'ordre de grandeur* du pas de la table.

L'*interpolation linéaire* est une méthode plus précise de calcul de $F(x)$ qui consiste en ce qu'au lieu de (31) on construit une fonction linéaire par morceaux en utilisant les valeurs de gauche f_k et de droite f_{k+1} de F sur chaque intervalle (fig. 4). Cette fonction est de la forme

$$P_1(x) = f_k + \frac{f_{k+1} - f_k}{x_{k+1} - x_k}(x - x_k), \quad x_k \leqslant x \leqslant x_{k+1}. \qquad (33)$$

On peut effectuer l'interpolation par fonctions quadratiques, cubiques, etc., en utilisant à cet effet trois, quatre, etc., points de la table. Considérons le cas général.

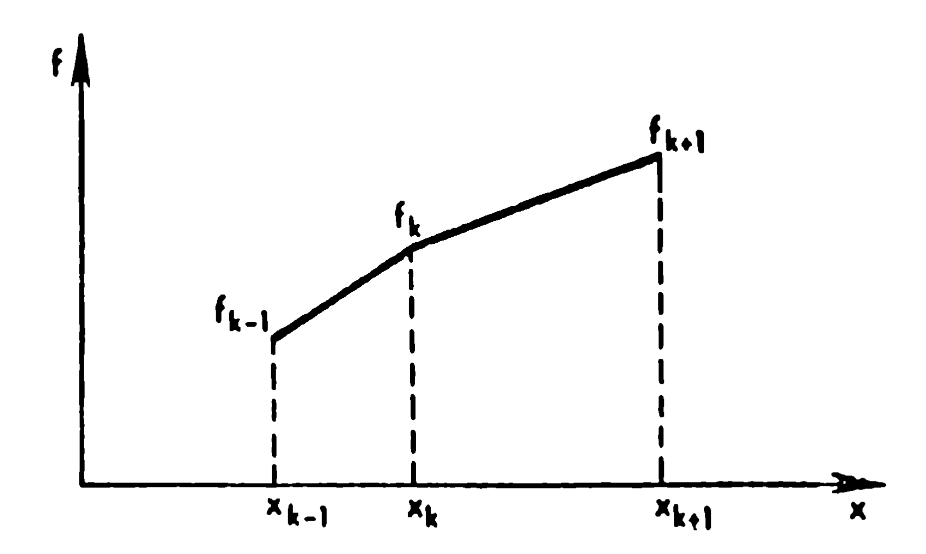


Fig. 4

Soient $n+1$ points $x_0, x_1, \ldots, x_n$ appelés *nœuds de l'interpolation* et les valeurs correspondantes $f_0, f_1, \ldots, f_n$. Formons le polynôme d'ordre n

$$P_n(x) = \sum_{m=0}^{n} a_m x^m, \tag{34}$$

dont les valeurs aux points x_i ($i = 0, 1, \ldots, n$) sont égales à f_i. Posant dans (34) $x = x_i$ on obtient un système de $n+1$ équations permettant de trouver $n+1$ inconnues, à savoir les coefficients du polynôme

$$\sum_{m=0}^{n} a_m x_i^m = f_i, \quad i = 0, 1, \ldots, n. \tag{35}$$

Le déterminant de ce système linéaire $|x_i^m|$ (déterminant de Vandermonde) est égal à $\prod_{i,j}(x_i - x_j)$, c'est-à-dire il est différent de zéro car tous les x_i sont différents. Par conséquent, le système (35) a une solution unique : un ensemble de a_m. Nous avons démontré qu'il existe un seul polynôme d'ordre n (au plus) prenant en $(n+1)$ points des valeurs données. Ce polynôme est appelé *polynôme d'interpolation*.

Il peut s'écrire différemment. La forme (34) est rarement utilisée car les expressions des coefficients a_m en fonction de x_i, f_i sont compliquées. Nous allons écrire le polynôme d'interpolation sous la *forme de Lagrange*, à savoir

$$P_n(x) = \sum_{i=0}^{n} f_i \frac{q_i(x)}{q_i(x_i)}, \tag{36}$$

où

$$q_i(x) = (x - x_0)(x - x_1) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n).$$

Comme $q_i(x_k) = 0$ pour $i \neq k$ on a de toute évidence $P_n(x_k) = f_k$. D'un autre coté, chaque $q_i(x)$ est un polynôme de degré n. Par conséquent leur combinaison linéaire (36) est le polynôme d'interpolation.

Pour $n = 1$ la formule (36) devient

$$P_1(x) = f_0 \frac{x - x_1}{x_0 - x_1} + f_1 \frac{x - x_0}{x_1 - x_0}, \tag{37}$$

c'est la formule d'interpolation linéaire (comparer avec (33)).

Nous allons estimer la précision avec laquelle le polynôme d'interpolation reproduit la fonction $F(x)$. La différence $P_n(x) - F(x)$ étant nulle dans les nœuds d'interpolation $x_0, x_1, \ldots, x_n$, le quotient de cette différence par la fonction

$$q(x) = (x - x_0)(x - x_1) \ldots (x - x_n) \tag{38}$$

est une fonction bornée et l'on peut écrire

$$F(x) = P_n(x) + R(x)\,q(x). \tag{39}$$

Nous allons estimer $R(x)$. A cet effet considérons la fonction auxiliaire

$$v(\xi) = F(\xi) - P_n(\xi) - R(x)\,q(\xi) = (R(\xi) - R(x))\,q(\xi). \tag{40}$$

Il est évident que la fonction $v(\xi)$ s'annule au moins en $n + 2$ points $x_0, x_1, \ldots, x_n, x$. Par conséquent, on aura

au moins un point $\xi(x)$ où la $(n+1)$-ième dérivée de $v(\xi)$ s'annule, si évidemment cette dérivée existe et qu'elle soit continue. En dérivant (40) $n+1$ fois et en substituant $\xi = \xi(x)$, on obtient :

$$0 = F^{(n+1)}(\xi(x)) - R(x)(n+1)!,$$

la $(n+1)$-ième dérivée du polynôme $P_n(\xi)$ de degré n étant nulle et $q(\xi)$ étant un polynôme de la forme $\xi^{n+1} + \ldots$. Ainsi

$$R(x) = \frac{F^{(n+1)}(\xi(x))}{(n+1)!},$$

on obtient ainsi l'expression de l'*erreur d'interpolation*, à savoir

$$F(x) - P_n(x) = \frac{1}{(n+1)!} F^{(n+1)}(\xi(x))\, q(x). \qquad (41)$$

La fonction $\xi(x)$ reste évidemment indéterminée.

Si le pas de la table n'est pas supérieur à un certain h, c'est-à-dire si

$$x_{k+1} - x_k \leqslant h, \quad k = 0, 1, \ldots, n-1,$$

on a $q(x) \sim h^{n+1}$ et on obtient à partir de (41)

$$|F(x) - P_n(x)| \leqslant c_n \max_x |F^{(n+1)}(x)|\, h^{n+1}, \qquad (42)$$

où c_n est une certaine constante. On en conclut que pour $h \to 0$ l'erreur décroît comme h^{n+1}.

La relation existant entre la valeur de l'erreur et le degré n du polynôme d'interpolation est plus compliquée. Dans la pratique courante on utilise très rarement des formules d'interpolation de degré assez grand. Ceci pour les raisons suivantes. Tout d'abord comme on peut le voir à partir de (42), l'augmentation du degré du polynôme conduit à une diminution de l'erreur d'interpolation seulement pour des fonctions très *lisses* ayant un grand nombre de dérivées. Mais généralement nous ne disposons pas d'une

information aussi grande sur les propriétés de la fonction $F(x)$. Ensuite les valeurs f_k sont toujours des valeurs approchées de $F(x_k)$ ne serait-ce que par arrondissement. C'est pourquoi les polynômes obtenus à partir de f_k et $F(x_k)$ différeront au plus d'une grandeur de l'ordre de $f_k - F(x_k)$. De plus les erreurs contenues dans f auront toujours un caractère aléatoire, ce qui montre que les fonctions les représentant ne sont pas lisses.

La méthode décrite de restitution de la fonction $F(x)$ à l'aide de la table x_k, f_k à partir du polynôme d'interpolation n'est pas la seule possible. Nous avons obtenu $P_n(x)$ (34) à l'aide d'un système de polynômes x^m ($m = 0, 1, \ldots$). Mais à cette fin on peut utiliser de nombreux autres systèmes $\varphi_m(x)$. Dans ce cas au lieu de (34) il est nécessaire d'envisager

$$\Phi_n(x) = \sum_{m=0}^{n} a_m \varphi_m(x) \tag{43}$$

et d'étudier les possibilités et la qualité de l'approximation. On peut aller plus loin et trouver la fonction approximante sous la forme d'une combinaison non linéaire de fonctions de base $\varphi_m(x)$. Mais il est évident qu'il faut avoir pour cela des raisons suffisantes car l'avantage ne peut être que dans le rétrécissement du domaine d'application possible de la méthode et dans l'utilisation d'une information essentielle supplémentaire sur $F(x)$ en plus de la table des valeurs.

On peut aborder différemment la question de la restitution d'une fonction d'après un ensemble de ses valeurs. Lors de la recherche de la fonction d'interpolation nous avons exigé que ses valeurs dans les nœuds x_k coïncident exactement avec f_k. Mais souvent il suffit d'exiger un écart minimal de ces valeurs à celles tabulées. Cette méthode est justifiée par exemple dans le cas où f contient à priori des erreurs importantes ou si seule la forme de la fonction approximante nous intéresse et non la précision de l'approximation.

Nous allons décrire une des méthodes de ce type, à savoir la *méthode des moindres carrés.* Soit la table x_k, f_k ($k = 0, 1, \ldots, n$). Il y a lieu de trouver la fonction

$$\Phi_M(x) = \sum_{m=0}^{M} a_m \varphi_m(x), \tag{44}$$

où $\varphi_m(x)$ est un système donné de fonctions (par exemple $\varphi_m(x) = x^m$) de telle sorte que la grandeur

$$\sum_{k=0}^{n} (\Phi_M(x_k) - f_k)^2 = \delta \tag{45}$$

soit minimale. Si $M \geqslant n$, on peut résoudre le problème en trouvant la fonction d'interpolation (44) pour laquelle $\delta = 0$; ce qui nous intéresse, c'est le cas où $M < n$.

Le choix des coefficients $a_0, a_1, \ldots, a_M$ est arbitraire, donc δ est fonction de ces coefficients. Pour trouver le minimum de la fonction $\delta(a_0, a_1, \ldots, a_M)$, annulons les dérivées de cette fonction par rapport à $a_0, a_1, \ldots, a_M$, on obtient ainsi le système d'équations

$$\sum_{l=0}^{M} a_l \sum_{k=0}^{n} \varphi_m(x_k)\varphi_l(x_k) = \sum_{k=0}^{n} \varphi_m(x_k) f_k, \quad m = 0, 1, \ldots, M.$$

En résolvant ce système d'équations linéaires nous trouvons les valeurs $a_0, a_1, \ldots, a_M$ déterminant complètement la fonction $\Phi_M(x)$ de (44) qui se trouve être la meilleure parmi les fonctions de cette forme pour l'approximation de la table x_k, f_k, si (45) est la mesure de l'écart.

Nous allons nous arrêter sur certaines applications des résultats obtenus. Telle ou telle méthode de correspondance entre la table et la fonction permet de réaliser à partir de la table différentes opérations fonctionnelles, à savoir l'intégration et la dérivation.

Ainsi, s'il faut calculer l'intégrale d'une fonction donnée par une table x_k, f_k ($k = 0, 1, \ldots, K$), en utilisant sur chaque intervalle (x_k, x_{k+1}) la formule d'interpolation

linéaire (33) on a

$$\int_{x_k}^{x_{k+1}} P_1(x)\,dx = \frac{f_k + f_{k+1}}{2}(x_{k+1} - x_k). \tag{46}$$

En prenant la somme de ces expressions sur tous les intervalles, on obtient la méthode de calcul de l'intégrale. Dans le cas où le pas du tableau est constant ($x_{k+1} - x_k = h$) on a

$$\int_{x_0}^{x_K} P_1(x)\,dx = \left(\frac{1}{2} f_0 + f_1 + f_2 + \ldots + f_{K-1} + \frac{1}{2} f_K\right) h, \tag{47}$$

c'est là la *formule quadratique des trapèzes* bien connue.

On peut estimer la précision avec laquelle la formule (47) donne la grandeur de l'intégrale de la fonction $F(x)$. Vu (41) pour $n = 1$ l'erreur d'interpolation est

$$\varepsilon(x) = \frac{1}{2} F''(\xi(x))(x - x_k)(x - x_{k+1}), \quad x_k \leqslant x \leqslant x_{k+1}, \tag{48}$$

d'où

$$\left|\int_{x_k}^{x_{k+1}} \varepsilon(x)\,dx\right| \leqslant \text{const} \max_{x_k \leqslant x \leqslant x_{k+1}} |F''|(x_{k+1} - x_k)^3.$$

En prenant la somme de cette inégalité sur tous les intervalles compte tenu de $Kh = x_K - x_0$, on obtient

$$\left|\int_{x_0}^{x_K} \varepsilon(x)\,dx\right| \leqslant \text{const} \max_{x_0 \leqslant x \leqslant x_K} |F''| h^2, \tag{49}$$

c'est-à-dire la précision de la formule des trapèzes (47) est *de l'ordre de* h^2. En utilisant d'autres fonctions d'interpolation $P_0(x)$, $P_2(x)$, etc., on obtient les formules quadratiques des rectangles, de Simpson, etc.

Le calcul de la dérivée d'une fonction tabulée est également très simple. En utilisant par exemple l'interpolation linéaire (33), on obtient

$$\frac{dF}{dx} \sim \frac{dP_1}{dx} = \frac{f_{k+1} - f_k}{x_{k+1} - x_k}, \qquad x_k \leqslant x \leqslant x_{k+1}. \tag{50}$$

En dérivant l'expression de l'erreur d'interpolation (48) on a

$$\frac{d\varepsilon}{dx} = \frac{1}{2} F''' (\xi) (x)) \xi' (x) (x - x_k) (x - x_{k+1}) + \\ + \frac{1}{2} F'' (\xi (x)) (2x - x_k - x_{k+1}).$$

Le second terme dans le second membre est de l'ordre de $x_{k+1} - x_k$, c'est-à-dire de h. C'est la précision du calcul de la dérivée assurée par la formule (50), à l'exclusion du point central de l'intervalle $x = (x_k + x_{k+1})/2$. En ce point le second terme s'annule et par conséquent la formule (50) donne la valeur de la dérivée en ce point à h^2 près.

Lorsque l'on essaie de calculer la dérivée seconde par interpolation linéaire, on obtient :

$$\frac{d^2F}{dx^2} \sim \frac{d^2P_1}{dx^2} = 0,$$

ce qui ne convient évidemment pas. Ceci se trouve en accord avec le fait que

$$\frac{d^2\varepsilon}{dx^2} = F'' (\xi (x)) + \ldots,$$

c'est-à-dire que l'erreur est finie, ne décroît pas lorsque h diminue. Pour le calcul des dérivées d'ordre plus élevé il y a lieu d'utiliser une interpolation d'ordre plus élevé.

Nous avons envisagé seulement le cas des fonctions d'une seule variable. Lorsque l'on passe à des problèmes multidimensionnels, les principes de base des méthodes exposées restent valables, mais de nombreux autres problèmes apparaissent.

Problèmes

1. Trouver c_n en fonction de n dans la formule (42).
2. Soit la table

x	0	1	2
f	0	1	1

Trouver l'approximation de cette table par une fonction linéaire en utilisant la méthode des moindres carrés. Puis trouver son approximation par une fonction linéaire en utilisant pour l'estimation de l'écart non pas (45) mais

$$\max_h |\Phi(x_h) - f_k| = \delta.$$

Comparer les fonctions obtenues avec le polynôme d'interpolation du second degré obtenu à partir de ces mêmes points.

3. Trouver les formules quadratiques en utilisant les polynômes d'interpolation de degrés zéro et deux. Trouver l'estimation aussi précise que possible de l'ordre de grandeur de l'erreur de ces formules.

4. Trouver la formule de calcul de la dérivée seconde d'une fonction tabulée. Estimer l'erreur.

5. Une fonction de deux variables $F(x, y)$ est donnée par la table suivante

x	0	h	0	$-h$	0
y	0	0	h	0	$-h$
F	$f_{0,0}$	$f_{1,0}$	$f_{0,1}$	$f_{-1,0}$	$f_{0,-1}$

Trouver la formule quadratique pour le calcul de l'intégrale

$$\iint\limits_{x^2+y^2\leq h^2} F(x, y)\,dx\,dy,$$

en utilisant dans chaque octante l'interpolation linéaire sur deux variables.

§ 3. ÉQUATIONS DIFFÉRENTIELLES ORDINAIRES

Nous allons étudier un problème typique, bien que pas très compliqué, pour les équations de ce genre. Il y a lieu de trouver la fonction $U(t)$ satisfaisant pour $t > 0$ à l'équation

$$\frac{dU}{dt} = F(t, U) \tag{51}$$

prenant pour $t = 0$ une valeur donnée

$$U(0) = U_0 \tag{52}$$

En théorie des équations différentielles ordinaires, si le second membre de (51), $F(t, U)$ satisfait, en tant que fonction de ses arguments, à certaines conditions de régularité, la solution du problème (51), (52), $U(t)$ existe, elle est unique et se trouve être une fonction régulière. Nous allons supposer que ces conditions soient vérifiées.

Les cas où $U(t)$ peut s'exprimer à l'aide de fonctions élémentaires ou par des intégrales de ces fonctions sont exceptionnels. En général, seules les méthodes numériques permettent de résoudre le problème (51), (52). Elles donnent évidemment une information limitée, approchée sur la solution mais sont en revanche universelles. Nous allons exposer ici la plus simple de ces méthodes, à savoir la *méthode d'Euler*.

On remplace avant tout le domaine de la variation continue de l'argument $t \geqslant 0$ par un ensemble discret de points

$$t_n = n\tau, \qquad n = 0, 1, 2, \ldots, \tag{53}$$

où τ est un certain nombre donné petit, dit le *paramètre de la méthode numérique*. Au lieu de la fonction $U(t)$ nous allons considérer la table

$$t_n, \ u_n, \ n = 0, 1, 2, \ldots \tag{54}$$

Comme par définition de la dérivée dU/dt est la limite du rapport $(U(t + \tau) - U(t))/\tau$ pour $\tau \to 0$, en remplaçant la dérivée par ce *rapport fini*, on obtient au lieu de l'équation différentielle (51) l'*équation aux différences*

$$\frac{u_{n+1} - u_n}{\tau} = F(t_n, u_n), \qquad n = 0, 1, 2, \ldots \tag{55}$$

ou

$$u_{n+1} = u_n + \tau F(t_n, u_n), \qquad n = 0, 1, 2, \ldots, \tag{55'}$$

Ainsi, compte tenu de (52) et posant

$$u_0 = U_0, \tag{56}$$

on peut à partir de (55') trouver successivement tous les u_n. On a ainsi trouvé l'algorithme de calcul.

Naturellement, maintenant l'algorithme est d'autant meilleur que le tableau t_n, u_n reproduit mieux la solution exacte du problème initial (51), (52), à savoir la fonction $U(t)$. Pour le voir, posons

$$u_n = U(t_n) + \delta u_n, \tag{57}$$

et essayons d'estimer la grandeur de l'erreur δu_n.

En substituant (57) dans l'équation aux différences (55), on obtient

$$\frac{\delta u_{n+1} - \delta u_n}{\tau} = F(t_n, U(t_n) + \delta u_n) - \frac{U(t_{n+1}) - U(t_n)}{\tau}. \tag{58}$$

Nous allons estimer le deuxième membre. Comme $U(t)$ est une fonction régulière et qu'elle satisfait à (51), on a

$$U(t_{n+1}) = U(t_n) + \tau \left(\frac{dU}{dt}\right)_{t_n} + O(\tau^2) = \\ = U(t_n) + \tau F(t_n, U(t_n)) + O(\tau^2),$$

d'où

$$\frac{U(t_{n+1}) - U(t_n)}{\tau} = F(t_n, U(t_n)) + O(\tau). \quad (59)$$

En comparant cette égalité avec (55), notons que la solution exacte de l'équation différentielle du problème initial satisfait à l'équation aux différences (55) à $O(\tau)$ près. En substituant maintenant (59) dans le deuxième membre de (58), on peut écrire

$$\frac{\delta u_{n+1} - \delta u_n}{\tau} = F(t_n, U(t_n) + \delta u_n) - F(t_n, U(t_n)) + O(\tau). \quad (60)$$

Vu les conditions de régularité imposées à $F(t, U)$, on a l'inégalité

$$|F(t_n, U(t_n) + \delta u_n) - F(t_n, U(t_n))| \leqslant M |\delta u_n|, \quad (61)$$

où M est une constante. En utilisant (61) on obtient à partir de (60)

$$|\delta u_{n+1}| \leqslant (1 + \tau M) |\delta u_n| + O(\tau^2). \quad (62)$$

La formule (62) donne l'accroissement de l'erreur pour un pas. Donc, pour l'erreur accumulée durant N pas, on a

$$\delta u_N | \leqslant O(\tau^2) + (1 + \tau M) |\delta u_{N-1}| \leqslant \\ \leqslant O(\tau^2) + (1 + \tau M)(O(\tau^2) + (1 + \tau M) |\delta u_{N-2}|) \leqslant \\ \cdots \\ \leqslant O(\tau^2) + (1 + \tau M) O(\tau^2) + (1 + \tau M)^2 O(\tau^2) + \\ \cdots \\ + (1 + \tau M)^{N-1} O(\tau^2) + (1 + \tau M)^N |\delta u_0|,$$

d'où, en prenant la somme de la progression géométrique à raison $1 + \tau M$, on a

$$|\delta u_N| \leqslant \frac{(1+\tau M)^N - 1}{(1+\tau M) - 1} O(\tau^2) + (1+\tau M)^N |\delta u_0|. \qquad (63)$$

Ce qui nous intéresse c'est la valeur de l'erreur pour des τ *petits* pour t *donné* quelconque. Posant $t = N\tau$ et tenant compte de ce que

$$(1+\tau M)^N = (1+\tau M)^{t/\tau} \sim e^{Mt} \qquad \text{pour} \quad \tau \sim 0,$$

on obtient à partir de (63)

$$|\delta u(t)| \leqslant e^{Mt} O(\tau) + e^{Mt} |\delta u_0|. \qquad (64)$$

Comme $\delta u_0 = u_0 - U_0 = 0$, il en résulte que pour $\tau \to 0$ l'erreur δu sur un intervalle fini quelconque tend vers zéro.

Nous avons ainsi démontré que pour τ suffisamment petit la table obtenue par la méthode d'Euler donne une approximation aussi précise que l'on veut de la solution du problème initial (51), (52).

Nous n'avons pas obtenu le critère permettant de dire quel est τ que l'on peut considérer « suffisamment petit » pour obtenir une précision donnée. Nous n'avons que simplement établi la *convergence* de la solution approchée vers la solution exacte pour $\tau \to 0$.

On aurait pu en utilisant au lieu des désignations $O(\tau)$, $O(\tau^2)$ des expressions plus détaillées donner une estimation quantitative de la grandeur de l'erreur. A cet effet on n'aurait eu qu'à surmonter des dificultés purement techniques. Mais une telle estimation serait très surélevée et ne saurait être utilisée pour calculer l'erreur exacte.

C'est pourquoi, en cas de besoin, pour trouver la précision accessible, on procède comme suit. Comme $u \to U$ pour $\tau \to 0$, à partir d'un certain τ les premières décimales significatives u cessent de changer, elles correspondent donc à la solution exacte. En effectuant les calculs avec différents τ

et en comparant les résultats obtenus, on peut se faire une idée de la précision obtenue.

De ce point de vue les méthodes accusant une forte dépendance de la valeur de l'erreur du paramètre τ, i.e. les méthodes *de haute précision*, sont plus commodes. La méthode d'Euler a une précision d'ordre minimal, $u - U = O(\tau)$. Il est évident que ceci provient de l'approximation grossière employée consistant à remplacer l'équation différentielle par une équation aux différences.

Pour estimer la *qualité de l'approximation* on détermine la précision avec laquelle la solution du problème initial satisfait à l'équation aux différences. En comparant (55) et (59), on voit que dans le cas présent l'approximation est de l'ordre de $O(\tau)$. Cette méthode d'estimation de l'ordre de l'approximation est quelque peu arbitraire. Par exemple si au lieu de (55) on utilise (55′), l'ordre de l'approximation sera $O(\tau^2)$, vu la norme différente de l'équation aux différences (multiplication par τ). Nous allons utiliser la norme naturelle de l'équation aux différences pour laquelle cette dernière devient, à la limite, l'équation différentielle initiale. Ainsi, (55) (plus exactement (59)), pour $\tau \to 0$, devient (51), alors que (55′) devient l'égalité $U = U$ où les propriétés du problème initial n'apparaissent pas du tout.

Considérons l'équation aux différences

$$\frac{u_{n+1} - u_n}{\tau} = \frac{1}{2}\left(F(t_n, u_n) + F(t_{n+1}, u_{n+1})\right) \tag{65}$$

et estimons l'ordre de l'approximation de l'équation différentielle (51) correspondante. En substituant dans (65) au lieu de u_n, u_{n+1} les valeurs $U(t_n)$, $U(t_{n+1})$ et en utilisant les développements

$$U(t_{n+1}) = U(t_n) + \tau\left(\frac{dU}{dt}\right)_{t_n} + \frac{\tau^2}{2}\left(\frac{d^2U}{dt^2}\right)_{t_n} + O(\tau^3),$$

$$F(t_{n+1}, U(t_{n+1})) = F(t_n, U(t_n)) + \tau\left(\frac{dF}{dt}\right)_{t_n} + O(\tau^2),$$

on obtient l'égalité

$$\left(\frac{dU}{dt}+\frac{\tau}{2}\frac{d^2U}{dt^2}\right)_{t_n}=\left(F+\frac{\tau}{2}\frac{dF}{dt}\right)_{t_n}+O(\tau^2).$$

Comme $U(t)$ satisfait à l'équation différentielle (51), ainsi qu'à

$$\frac{d^2U}{dt^2}=\frac{dF}{dt}$$

en découlant, on en conclut que pour $u = U$ (65) est satisfait à $O(\tau^2)$ près.

Ainsi l'équation aux différences (65) est une approximation de l'équation initiale (51) à $O(\tau^2)$ près. Le fait que la solution conserve cette même précision, c'est-à-dire que $u - U = O(\tau^2)$ n'est pas évident, et il y a lieu de démontrer cette propriété. Mais nous n'allons pas nous arrêter sur ce point.

A la différence de (55), la formule aux différences (65) ne permet pas en général d'exprimer d'une manière explicite u_{n+1} en fonction de u_n. C'est une équation par rapport à u_{n+1}. Pour trouver la solution, on peut utiliser tel ou tel processus d'itérations, d'autant plus que l'on a toujours une bonne approximation initiale u_n. Cependant il est inutile d'essayer de trouver la valeur exacte de u_{n+1}, ce serait pure perte de temps, car l'équation elle-même est donnée à $O(\tau^2)$ près. C'est pourquoi on peut se limiter aux deux itérations suivantes. Tout d'abord on calcule la première approximation $\tilde{u}_{n+1}$ par la formule de la méthode d'Euler

$$\tilde{u}_{n+1}=u_n+\tau F(t_n,\ u_n) \tag{66}$$

que l'on substitue ensuite dans le deuxième membre de (65). On trouve ainsi la valeur définitive de u_{n+1},

$$u_{n+1}=u_n+\frac{\tau}{2}(F(t_n,\ u_n)+F(t_{n+1},\ \tilde{u}_{n+1})). \tag{67}$$

En fait, ceci signifie qu'au lieu de (65) on utilise l'équation aux différences

$$\frac{u_{n+1}-u_n}{\tau}=\frac{1}{2}\left(F(t_n,\ u_n)+F(t_{n+1},\ u_n+\tau F(t_n,\ u_n))\right). \quad (68)$$

Il est facile de voir que cette équation, tout comme (65), est une approximation de l'équation initiale (51) à $O(\tau^2)$ près, donnant l'expression explicite de u_{n+1} en fonction de u_n.

Remarquons que lorsque l'on effectue le calcul à l'aide de (66), (67), on peut contrôler la précision du résultat obtenu en comparant les valeurs $\tilde{u}_{n+1}$ et u_{n+1} sur chaque pas de calcul. Une trop grande ou une trop petite valeur de la différence $u_{n+1}-\tilde{u}_{n+1}$ signifie qu'il y a lieu de diminuer ou d'augmenter le pas τ. En changeant le pas dans le sens convenable, on peut maintenir la précision à un niveau donné.

En généralisant la méthode exposée de l'équation aux différences on peut obtenir des méthodes dont la précision est d'un ordre encore plus élevé. Si l'on prend en considération plusieurs valeurs intermédiaires de $\tilde{u}$ calculées successivement, on arrive à des méthodes du type de celle de Runge-Kutta (voir problème 2). Une autre généralisation consiste à utiliser pour obtenir u_{n+1} non seulement u_n mais également les valeurs connues u_{n-1}, u_{n-2}, . . . Le principe général de ces méthodes (du type d'Adams) est le suivant. A partir de la suite donnée $F(t_n,\ u_n)$, $F(t_{n-1},\ u_{n-1})$, . . ., $F(t_{n-k},\ u_{n-k})$ on trouve le polynôme d'interpolation $P_k(t)$. En l'utilisant au lieu de F sur l'intervalle t_n, t_{n+1} (extrapolation), on peut écrire en intégrant (51)

$$u_{n+1}-u_n=\int_{t_n}^{t_{n+1}} P_k(t)\,dt.$$

Comme $P_k(t)$ est une fonction linéaire des valeurs de F aux points mentionnés, après intégration on obtient

$$u_{n+1} = u_n + \tau \sum_{i=0}^{k} c_i F(t_{n-i}, u_{n-i}), \tag{69}$$

c'est-à-dire une formule aux différences. Pour $k = 0$ (69) devient la formule de la méthode d'Euler (55'). Connaissant u_{n+1} on peut, comme dans la méthode d'Euler, introduire une correction en utilisant le polynôme d'interpolation tenant compte de la valeur u_{n+1} qui vient d'être obtenue. Cette méthode permet de réaliser des méthodes numériques d'une précision quelconque et ceci d'une manière très économique car l'augmentation de la précision s'obtient en utilisant les valeurs de u dont on dispose déjà.

Néanmoins ces méthodes, très répandues lorsque les calculs étaient effectués à la main, ne sont presque pas utilisées à l'heure actuelle. Ceci pour des raisons très caractéristiques pour les méthodes modernes de calcul numérique. En effet, on peut utiliser la formule (69) seulement à partir d'un certain $n = k$. Par conséquent les valeurs $u_1, u_2, \ldots$ $\ldots, u_k$ doivent être obtenues différemment, par des formules spéciales. Des difficultés analogues apparaissent lorsque l'on change le pas de l'intégration. La nécessité d'étendre l'algorithme pour tenir compte de plusieurs cas exceptionnels rend ces méthodes peu utilisées.

Tous ce qui vient d'être dit ci-dessus peut, tout naturellement, être généralisé à un système d'équations différentielles (et par conséquent à des équations d'ordre plus élevé). En particulier, si U, F, u ne sont pas des grandeurs scalaires mais des grandeurs vectorielles (51) devient un système d'équations différentielles, et les formules aux différences (55) et autres décrivent les méthodes d'intégration de ce système. Les résultats de l'étude de la convergence de l'approximation restent également bien que l'étude elle-même se trouve être un peu plus compliquée.

Lors de l'élaboration des méthodes numériques nous avons systématiquement et essentiellement supposé la fonction $F(t, U)$ régulière. Si l'on renonce à cette hypothèse et que l'on suppose que la fonction $F(t, U)$ présente des singularités, on ne sera pas de toute évidence obligé d'élaborer une méthode spéciale de solution qu'au voisinage de ces singularités. Par exemple, si la fonction $F(t, U)$ présente un point singulier, du type indétermination 0/0, il y a lieu d'isoler les termes principaux $F(t, U)$, de trouver la solution analytique approchée au voisinage de ce point et de l'utiliser pour passer le point singulier, en juxtaposant à la table t_n, u_n obtenue par la méthode numérique.

Enfin, nous allons nous arrêter sur le rôle des erreurs d'arrondissement. Comme nous l'avons déjà noté, toute équation aux différences porte une erreur d'approximation. Si l'imprécision apparaissant sur un pas par suite de l'arrondissement ne surpasse pas la précision de l'approximation, cette relation reste vraie pour les erreurs accumulées dues à l'une et l'autre de ces causes. Par conséquent, si avec la diminution de τ on augmente la précision des calculs, la convergence se conserve. Considérons par exemple la méthode d'Euler. En fait, le processus de calcul nous donne non pas une suite u_n satisfaisant à (55') mais une certaine suite voisine $\tilde{u}_n$. Cette dernière est obtenue par la formule suivante

$$\tilde{u}_{n+1} = \tilde{u}_n + \tau F(t_n, \tilde{u}_n) + \delta_n,$$

où δ_n est l'erreur admise lors du calcul du second membre. Elle est due à l'imprécision du calcul de $F(t_n, \tilde{u}_n)$ et à l'imprécision de la multiplication par τ. Il est évident que δ_n caractérise la précision des calculs. Si $\delta_n = 0(\tau^2)$, alors $\tilde{u}_n$ satisfait à la même équation que $U(t_n)$, c'est-à-dire à (59). Par conséquent, $\tilde{u}_n - u_n = O(\tau)$ et le processus de calcul réel mené avec une précision excédentaire $\delta =$ $= 0(\tau^2)$ permet d'obtenir la solution à $O(\tau)$ près.

Problèmes

1. Montrer que la solution obtenue par la méthode d'Euler avec les corrections ((66), (67)) coïncide avec la solution exacte à $O(\tau^2)$ près.

2. La méthode de Runge-Kutta d'intégration de l'équation (51) est décrite par les formules suivantes:

$$u_{n+1} = u_n + \frac{\tau}{6}(f_1 + 2f_2 + 2f_3 + f_4),$$

où

$$f_1 = F(t_n, u_n),$$

$$f_2 = F\left(t_n + \frac{\tau}{2}, \quad u_n + \frac{\tau}{2} f_1\right),$$

$$f_3 = F\left(t_n + \frac{\tau}{2}, \quad u_n + \frac{\tau}{2} f_2\right),$$

$$f_4 = F(t_n + \tau, \; u_n + \tau f_3).$$

Trouver l'ordre de l'approximation de cette méthode.

3. Montrer que la méthode d'Euler pour un .système linéaire

$$\frac{dU_i}{dt} = \sum_{j=1}^{I} a_{ij} U_j, \qquad i = 1, 2, \ldots, I$$

donne la solution à $O(\tau)$ près.

4. Décrire l'algorithme de calcul donnant la solution du système

$$\frac{dx}{dt} = \frac{x}{t} + yf(t),$$

$$\frac{dy}{dt} = x + t,$$

satisfaisant aux données initiales $t = 0$, $x = 0$, $y = 0$ et prenant pour un certain $t = T$ des valeurs données $x = X$, $y = Y$. La fonction $f(t)$ est régulière et donnée par une table admettant une interpolation linéaire.

5. Pour la solution du problème

$$\frac{dU}{dt}+MU=0, \quad U(0)=1$$

utiliser les deux équations aux différences

$$\frac{u_{n+1}-u_n}{\tau}+Mu_n=0$$

et

$$\frac{u_{n+1}-u_n}{\tau}+Mu_{n+1}=0.$$

Estimer l'ordre de l'approximation dans les deux cas. Comparer les solutions approchées, obtenues pour τ fini, avec celle exacte. Lequel des deux cas est préférable pour M grand lorsque seul le caractère qualitatif de la solution exacte nous intéresse ?

Chapitre II

§ 4. ÉQUATIONS AUX DÉRIVÉES PARTIELLES

Tout ce qui va suivre sera consacré aux méthodes numériques de solution des équations différentielles aux *dérivées partielles*. Lorsque l'on passe d'une à deux et plus variables indépendantes, la diversité et la complexité des problèmes augmentent brusquement. N'ayant pas la possibilité de voir en détail tous les aspects et les méthodes de cette théorie en plein essor nous nous limiterons à l'étude de certaines questions de principe.

Commençons par le problème le plus simple. Dans le domaine

$$-\infty < x < \infty, \quad t \geqslant 0 \tag{70}$$

il y a lieu de trouver la fonction $U(x, t)$ satisfaisant pour $t > 0$ à l'équation différentielle

$$\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} = F(x, t) \tag{71}$$

et prenant pour $t = 0$ des valeurs données

$$U(x, 0) = \Phi(x). \tag{72}$$

La solution exacte de ce problème est donnée par la formule

$$U(x, t) = \Phi(x - at) + \int_0^t F(x - at + at', t')\,dt', \tag{73}$$

comme on peut facilement s'en rendre compte. Mais pour le moment ce problème ne nous intéresse qu'à titre d'exemple élémentaire, permettant de montrer la plupart des propriétés

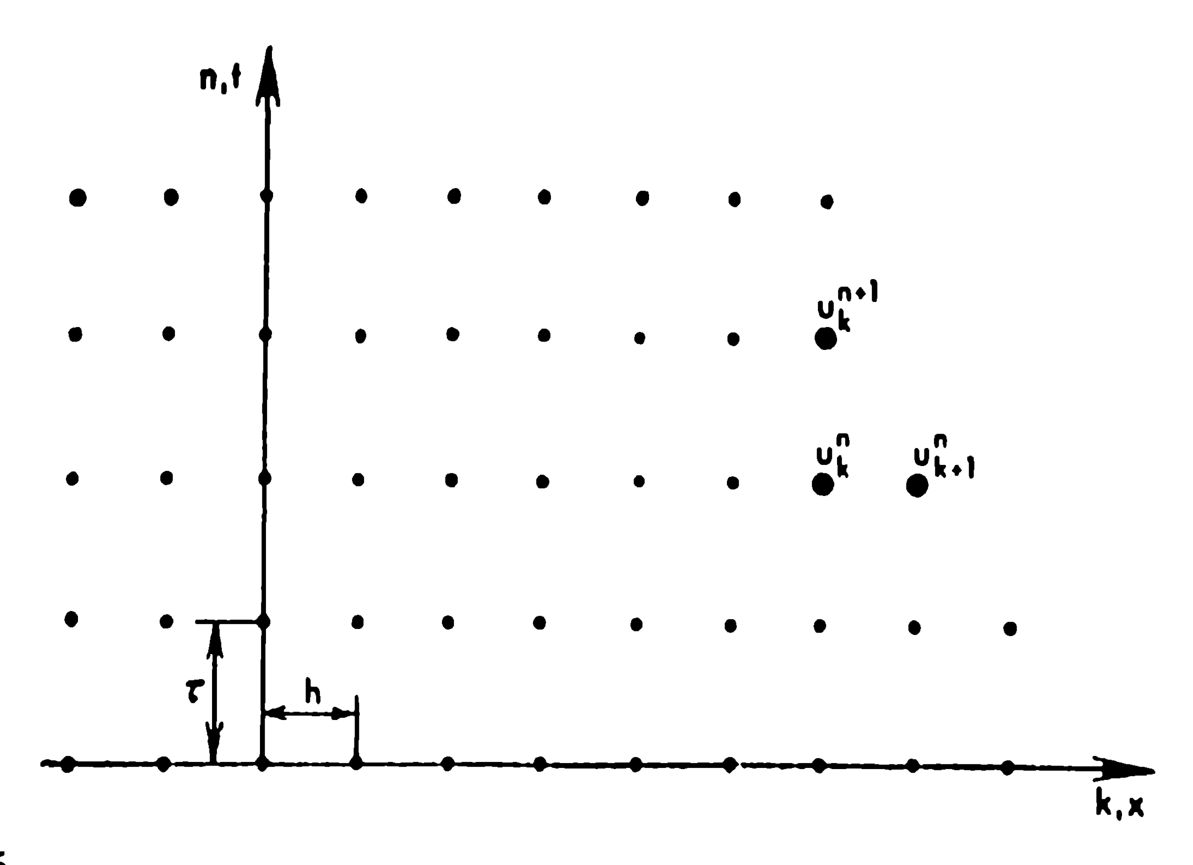


Fig. 5

des méthodes numériques d'intégration des équations différentielles aux dérivées partielles. Ci-dessous nous y reviendrons encore une fois.

Pour l'élaboration d'une méthode numérique de résolution du problème (70), (71), (72) il faut avant tout remplacer le domaine des variations continues des arguments (70) par un *réseau de calcul* (fig. 5), c'est-à-dire par un ensemble discre des points de coordonnées

$$\left.\begin{aligned} x_k &= kh, \qquad k = 0, \pm 1, \pm 2, \ldots \\ t^n &= n\tau, \qquad n = 0, 1, 2, \ldots \end{aligned}\right\} \tag{74}$$

On voit ainsi apparaître deux paramètres τ et h, ce sont les *pas du réseau*. Au lieu des fonctions $U(x, t)$, $F(x, t)$, $\Phi(x)$ on va envisager les *fonctions discrètes*, c'est-à-dire des suites numériques u_k^n, f_k^n, φ_k correspondant aux points du réseau x_k, t^n (74). Puis, en remplaçant les dérivées partielles figurant

dans (71) par les *rapports aux différences* on a

$$\frac{u_k^{n+1}-u_k^n}{\tau}+a\,\frac{u_{k+1}^n-u_k^n}{h}=f_k^n \tag{75}$$

pour chaque paire d'indices k, n, c'est-à-dire pour chaque point du réseau de calcul (74). Enfin, au lieu de (72) on écrit

$$u_k^0=\varphi_k. \tag{76}$$

Ainsi nous avons remplacé le problème (70), (71), (72) par le *problème numérique aux différences* (74), (75), (76). Il est évident que la méthode utilisée n'est pas unique. Le nombre de possibilités se présentant pour la résolution des équations aux dérivées partielles est bien plus grand que dans le cas des équations différentielles ordinaires. Mais pour le moment nous en resterons là.

L'algorithme de calcul donnant les valeurs de u_k^n est très simple. Nous allons résoudre (75) par rapport à u_k^{n+1},

$$u_k^{n+1}=\left(1+a\,\frac{\tau}{h}\right)u_k^n-a\,\frac{\tau}{h}\,u_{k+1}^n+\tau f_k^n. \tag{77}$$

Si les grandeurs u_k^n pour une certaine couche n de points sont données la formule (77) permet de les calculer pour la $(n+1)$-ième couche de points. Comme u_k^0 sont donnés, ils permettent de calculer u_k^1, puis u_k^2 et ainsi de suite.

Nous allons passer maintenant à la question essentielle, à savoir à quel point la solution u_k^n obtenue par une méthode numérique est voisine de la solution exacte du problème initial $U(x, t)$. Il est évident que ceci ne peut avoir lieu que si τ et h sont petits.

Posons

$$u_k^n=U(x_k,\ t^n)+\delta u_k^n \tag{78}$$

et substituons cette expression dans la formule aux différences (75). On obtient

$$\frac{\delta u_k^{n+1}-\delta u_k^n}{\tau}+a\frac{\delta u_{k+1}^n-\delta u_k^n}{h}=$$

$$=f_k^n-\left(\frac{U_k^{n+1}-U_k^n}{\tau}+a\frac{U_{k+1}^n-U_k^n}{h}\right). \qquad (79)$$

Ici et plus loin U_k^n désigne $U(x_k, t^n)$. Nous allons estimer le deuxième membre de (79). En supposant que $U(x, t)$ est une fonction régulière, pour τ et h petits on peut écrire

$$\left.\begin{aligned} U_k^{n+1}&=U_k^n+\tau\left(\frac{\partial U}{\partial t}\right)_k^n+O(\tau^2),\\ U_{k+1}^n&=U_k^n+h\left(\frac{\partial U}{\partial x}\right)_k^n+O(h^2),\end{aligned}\right\} \qquad (80)$$

d'où

$$\frac{U_k^{n+1}-U_k^n}{\tau}+a\frac{U_{k+1}^n-U_k^n}{h}=\left(\frac{\partial U}{\partial t}+a\frac{\partial U}{\partial x}\right)_k^n+O(\tau, h).$$

Comme $U(x, t)$ satisfait à (71) et $F(x_k, t^n)=f_k^n$, la dernière égalité signifie que

$$\frac{U_k^{n+1}-U_k^n}{\tau}+a\frac{U_{k+1}^n-U_k^n}{h}=f_k^n+O(\tau, h). \qquad (81)$$

En comparant (81) avec (75) on voit que la solution du problème initial $U(x, t)$ satisfait à l'équation aux différences (75) à $O(\tau, h)$ près, c'est-à-dire que nous avons une *approximation*.

En substituant (81) dans le deuxième membre de (79), on obtient l'équation déterminant δu

$$\frac{\delta u_k^{n+1}-\delta u_k^n}{\tau}+a\frac{\delta u_{k+1}^n-\delta u_k^n}{h}=O(\tau, h) \qquad (82)$$

ou

$$\delta u_k^{n+1}=\left(1+a\frac{\tau}{h}\right)\delta u_k^n-a\frac{\tau}{h}\delta u_{k+1}^n+\tau O(\tau, h). \qquad (83)$$

Envisageons tout d'abord le cas où a, τ, h satisfont aux inégalités

$$0 \leqslant -a \frac{\tau}{h} \leqslant 1. \tag{84}$$

Dans ce cas les coefficients près de δu_k^n et δu_{k+1}^n dans le deuxième membre de (83) sont positifs, et l'on peut écrire

$$|\delta u_k^{n+1}| \leqslant \left(1 + a \frac{\tau}{h}\right) |\delta u_k^n| + \left(-a \frac{\tau}{h}\right) |\delta u_{k+1}^n| + \tau O(\tau, h) \leqslant$$
$$\leqslant \max(|\delta u_k^n|, \ |\delta u_{k+1}^n|) + \tau O(\tau, h).$$

Introduisons la désignation

$$\| \delta u^n \| = \max_k |\delta u_k^n|, \tag{85}$$

l'inégalité précédente donne alors

$$\| \delta u^{n+1} \| \leqslant \| \delta u^n \| + \tau O(\tau, h), \tag{86}$$

c'est-à-dire que l'écart maximal δu pour un pas augmente de $\tau O(\tau, h)$ au plus. Pour N pas on aura

$$\| \delta u^N \| \leqslant \| \delta u^0 \| + N\tau O(\tau, h). \tag{87}$$

Fixons $t = N\tau$ fini quelconque et faisons tendre τ, h vers zéro, N tendra alors à l'infini. Comme $\delta u_k^0 = \Phi_k - \varphi_k$ on obtient à partir de (87)

$$\| \delta u(t) \| = O(\tau, h). \tag{88}$$

Nous avons ainsi démontré que si, lorsque τ, h tendent vers zéro, les conditions (84) se trouvent vérifiées, la solution du problème aux différences (74), (75), (76) tend vers la solution du problème initial (70), (71), (72).

Considérons maintenant le cas opposé lorsque τ, h tendent vers zéro de telle sorte qu'au moins l'une des conditions (84) ne soit pas vérifiée. Il se trouve que dans ce cas, en général, il n'y a pas de convergence. Nous allons montrer ceci par le raisonnement simple suivant.

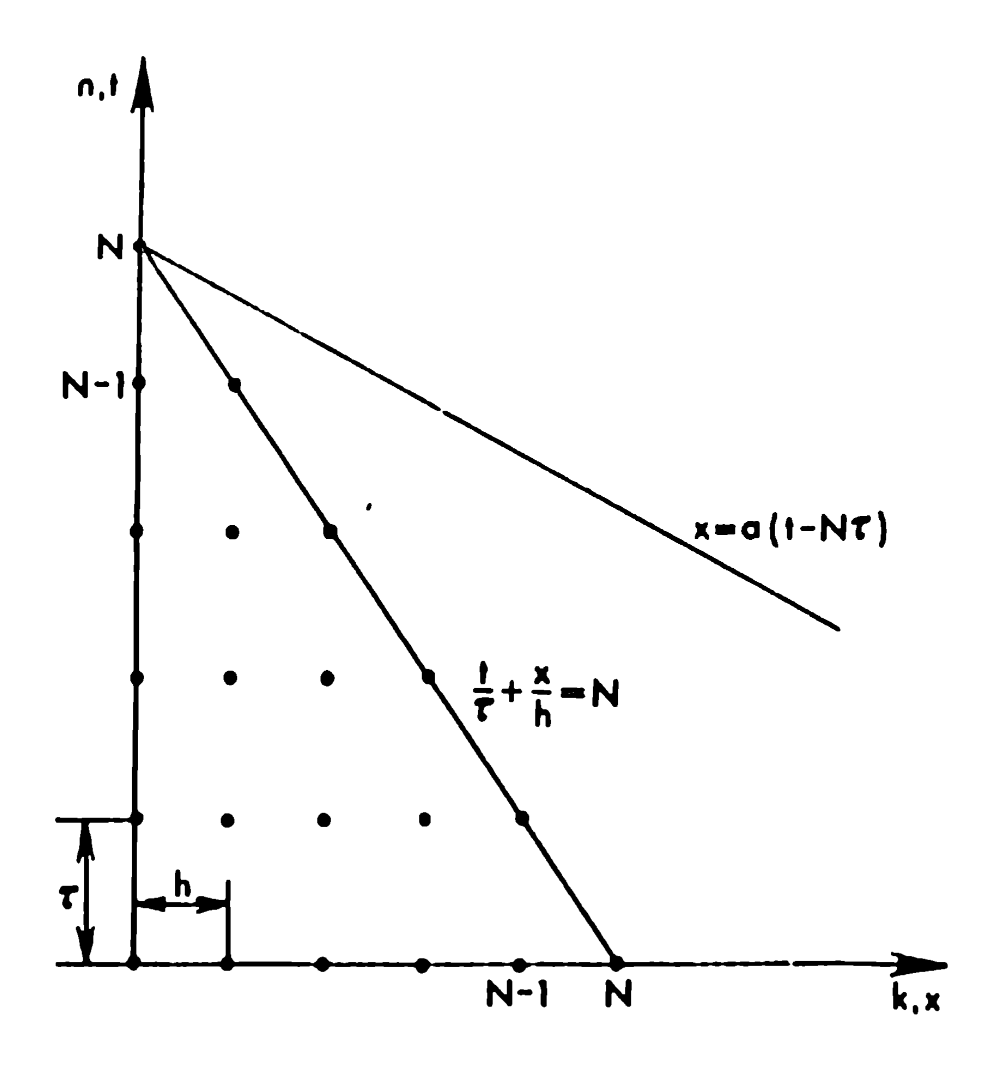


Fig. 6

Déterminons pour u_0^N le domaine de dépendance. Puisque u_0^N s'exprime en fonction de f_0^{N-1}, u_0^{N-1}, u_1^{N-1} qui s'expriment à leur tour en fonction de f_0^{N-2}, f_1^{N-2}, u_0^{N-2}, u_1^{N-2}, u_2^{N-2}, etc. (fig. 6), en prolongeant ce processus on peut exclure toutes les valeurs intermédiaires u_k^n et exprimer u_0^N directement en fonction de f_k^n et φ_k. Les points k, n, correspondant aux valeurs f et φ utilisées forment de toute évidence un triangle représenté fig. 6, qui se trouve être le domaine de dépendance de u_0^N. Ses côtés sont formés des segments de droites

$$x = 0, \quad t = 0, \quad \frac{t}{\tau} + \frac{x}{h} = N, \tag{89}$$

c'est-à-dire qu'il est determiné par deux paramètres $N\tau$ et τ/h. Si ces paramètres sont constants pour τ, $h \to 0$, alors le domaine de dépendance reste inchangé.

Nous allons maintenant déterminer le domaine de dépendance de la solution exacte en ce même point, soit $U(0, N\tau)$. Comme on peut le voir à partir de (73), $U(0, N\tau)$ se trouve entièrement déterminé par les valeurs $F(x, t)$ sur la droite

$$x = a(t - N\tau), \qquad 0 \leqslant t \leqslant N\tau, \tag{90}$$

et par la valeur Φ au point d'intersection de cette droite avec l'axe x.

Supposons que la droite (90) se trouve à l'extérieur du triangle (89). Dans ce cas $U(0, N\tau)$ et u_0^N seront déterminés par différents facteurs indépendants et nous n'avons aucune raison pour penser qu'il seront voisins. En changeant par exemple les valeurs F et Φ seulement au voisinage de la droite (90) nous obtiendrons différents $U(0, N\tau)$. De plus u_0^N ne sera pas influencé par ces variations.

Trouvons les conditions d'appartenance de la droite (90) au triangle (89). Pour $t < N\tau$ donné, on a pour les points intérieurs du triangle (89)

$$0 \leqslant x \leqslant h\left(N - \frac{t}{\tau}\right).$$

En substituant au lieu de x son expression (90) on obtient

$$0 \leqslant a(t - N\tau) \leqslant \frac{h}{\tau}(N\tau - t)$$

ou en simplifiant par $N\tau - t$

$$0 \leqslant -a \leqslant \frac{h}{\tau},$$

ce qui de toute évidence est identique à (84).

Ainsi nous avons montré que lorsque les conditions (84) ne sont pas vérifiées, en général, il n'y a pas de convergence.

Il y a lieu de faire la remarque importante suivante. Le problème aux différences (74), (75), (76) *donne une approximation* de l'équation différentielle initiale (70), (71), (72) que la condition (84) soit vérifiée ou non. Cependant il se trouve qu'une seule approximation ne suffit pas pour la

convergence de u_k^n vers la solution exacte $U(x, t)$. La condition (84) est, dans le problème présent, la condition supplémentaire assurant la convergence. Lorsque cette condition n'est pas vérifiée, l'écart de u_k^n à $U(x, t)$ peut être quelconque.

Néanmoins il est intéressant de voir le caractère de la solution u_k^n obtenue pour τ, h donnés et finis et ce qui se passe lorsque τ, h diminuent. Pour s'en rendre compte sans effectuer de calculs nous allons procéder comme suit. Revenons à la formule (83) décrivant le processus d'évolution de l'erreur d'une couche à l'autre. Vu la présence de $\tau O(\tau, h)$ dans le deuxième membre, l'ordre de l'erreur n'est pas inférieur à cette grandeur. Quant à l'erreur δu_k^n, c'est une certaine fonction compliquée de l'indice k. Supposons qu'on puisse l'écrire sous la forme d'une somme dont l'un des termes est de la forme $\varepsilon(-1)^k$ où de toute évidence ε est de l'ordre de $\tau O(\tau, h)$. Nous allons voir comment change cette seule composante de l'erreur, c'est-à-dire posons

$$\delta u_k^n = \varepsilon(-1)^k \tag{91}$$

et calculons δu_k^{n+1}, δu_k^{n+2}, . . . En idéalisant quelque peu le problème, nous ne tiendrons pas compte de l'influence du terme $\tau O(\tau, h)$ sur les pas suivants. On obtient

$$\delta u_k^{n+1} = \left(1 + a\frac{\tau}{h}\right)\varepsilon(-1)^k - a\frac{\tau}{h}\varepsilon(-1)^{k+1} =$$
$$= \left(1 + 2a\frac{\tau}{h}\right)\varepsilon(-1)^k,$$

autrement dit, la fonction de la forme (91) sur la couche suivante conserve sa valeur multipliée par $1 + 2a\tau/h$. Il est évident qu'après N couches ce facteur est égal à $(1 + 2a\tau/h)^N$, soit

$$\delta u_k^{n+N} = \left(1 + 2a\frac{\tau}{h}\right)^N \varepsilon(-1)^k, \tag{92}$$

et par conséquent l'évolution de la composante étudiée de l'erreur est donnée par la grandeur $1 + 2a\tau/h$. Si

$$\left|1+2a\frac{\tau}{h}\right|<1, \qquad (93)$$

l'erreur décroît, dans le cas contraire elle croît exponentiellement.

Ainsi la solution u_k^n contenant cette erreur perd rapidement son sens devenant une suite chaotique de termes très grands. Même si ε est petit, cela ne change rien et ne peut que reculer l'instant de la catastrophe. L'effet décrit est appelé *instabilité*.

Il est facile de voir que (93) est de nouveau la même condition (84). Mais ceci signifie que l'absence de convergence et l'instabilité sont dues à une même cause. En utilisant (92) pour un segment fini $t = N\tau$ on obtient pour le cas où la condition (93), ou (84), n'est pas vérifiée:

$$|\delta u_k^{n+1}|=\left|1+2a\frac{\tau}{h}\right|^{t/\tau}\tau O(\tau,\ h)\to\infty \quad \text{pour} \quad \tau,\ h\to 0.$$

Le calcul suivant le schéma aux différences instable non seulement ne donne pas un résultat voisin de la solution exacte mais est impossible. Dans ce cas la diminution de τ, h ne fait qu'accentuer cette instabilité.

Problèmes

1. Si $a > 0$, les conditions (84) ne sont jamais vérifiées quels que soient τ, h. Comment faut-il changer la formule aux différences (75) dans ce cas?

2. Elaborer et étudier la méthode aux différences de solution du problème (70), (71), (72) dans le cas général de la variable a, $a = a(x, t)$.

3. Pour résoudre le problème

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}, \quad U(x,\ 0)=\Phi(x)$$

envisager l'équation aux différences

$$\frac{u_k^{n+1} - u_k^n}{\tau} = \frac{u_{k+1}^n - 2u_k^n + u_{k-1}^n}{h^2}, \quad u_k^0 = \varphi_k.$$

Etudier son approximation (pour $2\tau/h^2 \leqslant 1$) et sa stabilité (sur la fonction $\varepsilon(-1)^k$).

§ 5. APPROXIMATION ET STABILITÉ

Le schéma de principe de l'étude de la convergence faite ci-dessus est typique pour un grand nombre de problèmes et peut être décrit de la manière suivante. Supposons qu'un certain problème de base différentiel soit remplacé par un problème aux différences du type

$$lu = f. \tag{94}$$

u, f sont ici des fonctions discrètes u_k^n, f_k^n et l un opérateur linéaire aux différences dépendant des paramètres τ, h. En particulier le problème envisagé (75), (76) s'écrit sous la forme (94) à condition de poser

$$lu = \begin{cases} \dfrac{u_k^{n+1} - u_k^n}{\tau} + a\,\dfrac{u_{k+1}^n - u_k^n}{h}, \\ u_k^0 \end{cases} \tag{95}$$

et

$$f = \begin{cases} f_k^n, \\ \varphi_k. \end{cases} \tag{96}$$

Pour l'étude de la convergence de la solution du problème (94) à la solution du problème initial, à savoir à la fonction U, on pose

$$u = U + \delta u$$

et en substituant cette expression dans (94) on obtient l'équation déterminant δu

$$l\delta u = f - lU. \tag{97}$$

Il est évident que la *convergence*, c'est-à-dire la tendance de δu vers 0, aura lieu si, premièrement, par le choix des paramètres τ, h le second membre de (97) peut être rendu aussi petit que l'on veut, et deuxièmement, s'il reste petit pour la solution de l'équation (97), c'est-à-dire pour la fonction discrète δu.

La première de ces conditions

$$lU - f \to 0 \quad \text{pour } \tau, h \to 0 \tag{98}$$

est la *condition d'approximation* que nous connaissons déjà (comparer avec (81)). Elle caractérise la relation existant entre le problème aux différences et le problème différentiel et établit le voisinage de ces problèmes.

La réalisation de la seconde condition dépend seulement des propriétés du problème aux différences, à savoir : l'opérateur aux différences l doit être tel que pour τ, h quelconques la solution du problème (97) soit du même ordre de grandeur que le deuxième membre, c'est-à-dire

$$\delta u \sim f - lU. \tag{99}$$

Peu d'opérateurs aux différences sont doués de ces propriétés. Comme nous l'avons montré ci-dessus pour l'opérateur l (95), la condition (99) est vraie si a, τ, h satisfont à la relation (84). Dans le cas contraire la condition (99) n'est pas vérifiée. Comme dans ce cas il y a instabilité, en attribuant à ce terme un sens plus large, la condition (99) est appelée *condition de stabilité*.

Remarquons que les équations (97) et (94) ne diffèrent que par les désignations : δu, $f - lU$ au lieu de u, f. La condition de stabilité (99) peut donc s'écrire sous la forme

$$u \sim f, \tag{100}$$

et la solution du problème (94) doit satisfaire à cette condition. Soulignons que l'opérateur l dépend des paramètres τ, h et la stabilité signifie que la relation (100) est vérifiée pour des τ, h aussi petits que l'on veut.

Pour donner un sens exact aux relations (98), (99), (100) il y a lieu de mentionner le mode d'estimation des grandeurs y figurant u, f, lU, etc., c'est-à-dire d'introduire une *norme* de ces fonctions de réseau $\| u \|$, $\| f \|$, $\| lU \|$, etc. Dans l'exemple envisagé ci-dessus nous avons utilisé en qualité de norme le maximum du module des valeurs de la fonction, autrement dit, l'égalité

$$\delta u = O(\tau, h)$$

signifie que

$$\| \delta u \| = \max_{k,\, n} | \delta u_k^n | = O(\tau, h).$$

D'autres définitions de la norme sont également possibles, ce qui importe, c'est qu'elle nous satisfasse en tant que méthode de mesure de la valeur exacte de la fonction de réseau. Ainsi la norme $\| \delta u \| = \max_{k\; n} h \; | \delta u_k^n |$ ne convient pas car, dans ce cas, $\| \delta u \| \to 0$ pour $h \to 0$ même si δu_k^n restent finis. Toute norme est une généralisation de la notion de valeur absolue d'un nombre et doit satisfaire aux relations

$$\| u + v \| \leqslant \| u \| + \| v \|, \qquad \| \alpha u \| = | \alpha | \; \| u \|,$$

où α est un nombre.

Nous n'avons supposé nulle part que les fonctions discrètes u, f figurant dans (94), (98), (100) sont des scalaires. Si le problème initial consiste dans l'intégration d'un système d'équations différentielles, dont la solution est un système de fonctions, c'est-à-dire un vecteur fonction, il est évident qu'alors u, f, lu sont des vecteurs fonctions discrètes. On doit en tenir compte lors de la définition d'une norme.

Ainsi pour un problème quelconque linéaire aux différences du type (94) la *convergence*

$$\|u - U\| \to 0 \quad \text{pour} \quad \tau,\ h \to 0 \tag{101}$$

découle de l'*approximation*

$$\|lU - f\| \to 0 \quad \text{pour} \quad \tau,\ h \to 0 \tag{102}$$

et de la *stabilité*

$$\|u\| \sim \|f\|.$$

Cette dernière relation signifie que, pour f quelconque, on a l'estimation suivante pour la solution correspondante u

$$\|u\| \leqslant \text{const}\, \|f\|, \tag{103}$$

où const ne dépend pas de τ et h.

Il est évident que si le problème est stable, (101) et (102) tendent vers zéro de la même manière, c'est-à-dire la précision de la solution et l'approximation sont du même ordre de grandeur.

La vérification de l'approximation, même dans le cas de problèmes compliqués, ne présente pas de difficultés. Vu la régularité de la solution exacte et compte tenu des équations différentielles auxquelles elle satisfait on peut calculer la grandeur $lU - f$ tout comme dans l'exemple simple envisagé ci-dessus.

Les particularités du problème jouent un rôle important dans l'étude de la stabilité. Il est rare que l'on puisse, comme dans l'exemple étudié, estimer directement u en fonction du deuxième membre f. Pour un grand nombre de problèmes il vaut mieux vérifier la stabilité à l'aide de solutions particulières, la méthode généralisant ayant permis dans l'exemple cité de faire apparaître l'instabilité. Nous reviendrons à cette question ci-dessous.

Passons aux problèmes non linéaires. Pour se faire une idée des difficultés apparaissant ici, nous allons commencer

par le problème suivant

$$\left.\begin{aligned}\frac{\partial U}{\partial t}+\left(\frac{\partial U}{\partial x}\right)^2=F(x,\ t),\\ U(x,\ 0)=\Phi(x).\end{aligned}\right\}\qquad(104)$$

Pour trouver la solution il est tout naturel d'utiliser le schéma aux différences

$$\left.\begin{aligned}\frac{u_k^{n+1}-u_k^n}{\tau}+\left(\frac{u_{k+1}^n-u_k^n}{h}\right)^2=f_k^n,\\ u_k^0=\varphi_k.\end{aligned}\right\}\qquad(105)$$

Ceci se base évidemment sur le fait que le problème (104) peut être approximé par le problème (105), la solution U du premier satisfaisant presque le second. En effet $U(x,\ t)$ étant lisse, en utilisant l'équation (104) on obtient

$$\left.\begin{aligned}\frac{U_k^{n+1}-U_k^n}{\tau}+\left(\frac{U_{k+1}^n-U_k^n}{h}\right)^2=f_k^n+O(\tau,\ h),\\ U_k^0=\varphi_k.\end{aligned}\right\}\qquad(106)$$

Pour estimer l'écart de u à U on substitue dans (105) $u=U+\delta u$ et compte tenu de (106) on obtient

$$\left.\begin{aligned}\frac{\delta u_k^{n+1}-\delta u_k^n}{\tau}+2\,\frac{U_{k+1}^n-U_k^n}{h}\cdot\frac{\delta u_{k+1}^n-\delta u_k^n}{h}+\\ +\left(\frac{\delta u_{k+1}^n-\delta u_k^n}{h}\right)^2=O(\tau,h),\\ \delta u_k^0=0.\end{aligned}\right\}\qquad(107)$$

La présence d'un terme quadratique ne permet pas d'estimer d'une manière efficace le comportement de δu. Supposons cependant qu'il y ait convergence et que δu soit petit, par exemple $\delta u \ll U$. Dans ce cas le comportement de δu sera déterminé par les termes linéaires principaux de (107),

c'est-à-dire par l'équation

$$\frac{\delta u_k^{n+1}-\delta u_k^n}{\tau}+2\,\frac{U_{k+1}^n-U_k^n}{h}\;\frac{\delta u_{k+1}^n-\delta u_k^n}{h}=O(\tau,\ h) \quad (108)$$

qui, à un coefficient près, coïncide avec l'équation envisagée antérieurement (82). Pour que cette coïncidence soit complète et que l'on puisse utiliser les résultats déjà obtenus, nous nous limiteront à l'étude de (108) au voisinage du point considéré x_k, t^n où l'on a

$$2\,\frac{U_{k+1}^n-U_k^n}{h}\sim 2\left(\frac{\partial U}{\partial x}\right)_k^n=a,$$

et (108) peut être remplacé par (82), soit:

$$\frac{\delta u_k^{n+1}-\delta u_k^n}{\tau}+a\,\frac{\delta u_{k+1}^n-\delta u_k^n}{h}=O(\tau,\ h).$$

Comme nous l'avons déjà établi, la solution de cette équation ne reste une grandeur petite qu'à condition que (84) soit vérifiée, celle-ci signifiant pour le moment que

$$-1\leqslant 2\,\frac{U_{k+1}^n-U_k^n}{h}\,\frac{\tau}{h}\leqslant 0. \quad (109)$$

Ainsi, pour que l'hypothèse de la convergence ne donne pas de contradiction il faut que la condition (109) soit vérifiée.

On ne peut affirmer que la question de la convergence pour le problème (105) soit complètement résolue, néanmoins la condition (109) donne beaucoup. En particulier, si la solution exacte a, dans un certain domaine, une dérivée positive $\frac{\partial U}{\partial x}>0$, les équations aux différences ne peuvent être utilisées. Pour vérifier la condition (109) au lieu des valeurs U, qui sont inconnues, on doit de toute évidence prendre u_k^n obtenues dans le calcul.

Appliquons les raisonnements faits au cas général d'un problème non linéaire aux différences que l'on peut

écrire

$$\mathscr{L}(u) = 0. \tag{110}$$

L'opérateur $\mathscr{L}$ étant non linéaire, il n'y a pas lieu d'isoler le deuxième membre f.

Avant tout on vérifie l'approximation, c'est-à-dire que la condition

$$\mathscr{L}(U) \to 0 \quad \text{pour } \tau, h \to 0 \tag{111}$$

soit vérifiée. Si cette condition n'est pas vérifiée rien ne donne d'espérer qu'il y aura convergence.

On aurait pu généraliser la notion de stabilité aux problèmes non linéaires, en définissant celle-ci à partir de la relation

$$u - U \sim \mathscr{L}(u) - \mathscr{L}(U). \tag{112}$$

Cette dernière ne pouvant être vérifiée, ce n'est pas la peine, bien que pour la convergence c'est justement la relation (112) qui manque; en effet, vu l'approximation, la différence $\mathscr{L}(u) - \mathscr{L}(U)$ est petite.

L'opérateur $\mathscr{L}$ est non linéaire et ses propriétés lorsqu'il est appliqué à différentes fonctions peuvent varier. Naturellement, ce qui nous intéresse en premier lieu, ce sont des fonctions voisines de la solution exacte. Si, ici déjà, les propriétés de $\mathscr{L}$ ne sont pas satisfaisantes, il n'y a pas lieu de s'attendre à la convergence. Mais pour $u - U = \delta u$ petits on peut obtenir l'information nécessaire sur $\mathscr{L}(u)$ en considérant sa *partie linéaire principale*, c'est-à-dire

$$\mathscr{L}(U) + \mathscr{L}'(U)\,\delta u, \tag{113}$$

où $\mathscr{L}'(U)$ est un opérateur linéaire (variante de $\mathscr{L}$) appliqué à δu et dépendant de U comme d'un paramètre. Remarquons que nous utilisons ici le fait que δu est petit, mais du point de vue formel ça n'a rien à voir avec la petitesse de τ, h.

Ainsi, en substituant dans (110) $u = U + \delta u$, en remplaçant $\mathscr{L}(U + \delta u)$ par (113), compte tenu de (111), on

obtient

$$\mathscr{L}'(U)\,\delta u \to 0 \quad \text{pour } \tau, h \to 0. \qquad (114)$$

Si ce problème provenant de l'opérateur linéaire $\mathscr{L}'(U)$ est stable, on peut s'attendre à ce qu'il y a convergence. En cas d'instabilité, la convergence n'a pas lieu.

En considérant l'opérateur $\mathscr{L}'(U)$ *localement*, dans un petit domaine du plan x, t, où $U(x, t)$ varie peu, on peut encore simplifier le problème en le ramenant à l'étude de la stabilité d'un problème linéaire à coefficients constants. En simplifiant, en simulant le problème il est important de ne pas omettre aucun de ses traits essentiels. Il faut une certaine habileté pour tenir compte au maximum du caractère spécifique du problème étudié.

Résumons donc. Lors de l'étude de la convergence d'un problème aux différences il y a lieu, avant tout, de vérifier les conditions d'approximation. En cas d'un résultat positif on procède à la *linéarisation* du problème, puis l'on maintient les coefficients constants, ce qui nous ramène à l'analyse de la stabilité d'un schéma linéaire aux différences à coefficients constants. Brièvement cette simplification, cette simulation du problème peut être représentée par la formule suivante

$$\mathscr{L}(u) \to \mathscr{L}'(U)\,\delta u \to l\delta u. \qquad (115)$$

Problèmes

1. Démontrer, par estimation directe, que le problème (108) est stable lorsque la condition (109) est vérifiée.

2. Trouver les conditions de stabilité (de convergence) pour les schémas aux différences approximant les problèmes

$$\frac{\partial U}{\partial t} = U\frac{\partial^2 U}{\partial x^2}, \qquad U(0, x) = U_0(x)$$

et

$$\frac{dU}{dt} = F(t, U), \qquad U(0) = U_0.$$

3. Elaborer les schémas aux différences approximant les problèmes

$$\frac{\partial U}{\partial t}=\frac{\partial}{\partial x}\mu(U)\frac{\partial U}{\partial x}, \quad U(0, x)=U_0(x)$$

et

$$\frac{\partial U}{\partial t}+\frac{\partial p(V)}{\partial x}=0, \quad U(0, x)=U_0(x),$$

$$\frac{\partial V}{\partial t}+\frac{\partial U}{\partial x}=0, \quad U(0, x)=V_0(x).$$

Effectuer la linéarisation des équations aux différences obtenues.

§ 6. CRITÈRE SPECTRAL DE STABILITÉ

Pour les problèmes linéaires aux différences du type $lu = f$ la stabilité signifie que $u \sim f$ c'est-à-dire que la solution et le second membre sont pour τ, $h \to 0$ du même ordre de grandeur. Nous allons nous limiter à l'étude des opérateurs l à structure en couches du type

$$lu=\begin{cases}\dfrac{u^{n+1}-Ru^n}{\tau}, & n=0, 1, \ldots,\\ u^0.\end{cases} \tag{116}$$

Ici u^n désigne la *fonction discrète sur n-ième couche*, c'est-à-dire l'ensemble u_k^n pour n donné, R étant un certain opérateur linéaire transposant une fonction sur une couche en une fonction sur une couche et dépendant des paramètres τ, h. Pour les opérateurs l du type (116) le problème $lu = f$ peut s'écrire sous la forme

$$\left.\begin{aligned}&u^{n+1}=Ru^n+\tau f^n, \quad n=0, 1, \ldots,\\ &u^0=v,\end{aligned}\right\} \tag{117}$$

où f^n, v sont des fonctions discrètes données sur les couches. C'est là la stratification de l'opérateur l à savoir les fonctions discrètes u^1, u^2, . . . peuvent être obtenues successivement à l'aide du même opérateur R, c'est-à-dire à l'aide de l'*opérateur de transition d'une couche à l'autre*.

Il y a lieu de trouver la relation existant entre les grandeurs u et f, v. Pour N quelconque les formules (117) permettent d'exprimer u^N en fonction de f^n ($n = N - 1$, $N - 2$, . . .) et v. En effet,

$$
\begin{aligned}
u^N &= \tau f^{N-1} + Ru^{N-1} = \\
&= \tau f^{N-1} + R(\tau f^{N-2} + Ru^{N-2}) = \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&= \tau(f^{N-1} + Rf^{N-2} + R^2 f^{N-3} + \ldots + R^{N-1} f^0) + R^N v = \\
&= \tau \sum_{m=0}^{N-1} R^m f^{N-m-1} + R^N v, \qquad (118)
\end{aligned}
$$

où R^m désigne la m-ième *puissance de l'opérateur* R et est lui aussi un opérateur linéaire.

Pour estimer les grandeurs des fonctions discrètes sur la couche v, u^n, f^n, Rv, . . ., introduisons une norme quelconque (par exemple $\| u^n \| = \max_k | u_k^n |$). Comme u^N s'exprime en fonction de f^n, et v au moyen de R^m, il est évident que la relation existant entre les grandeurs u^N et f^n, v dépendra des propriétés métriques des opérateurs R^m, autrement dit de la variation de la norme des fonctions discrètes lorsque l'on applique R^m. Supposons que les opérateurs R^m soient tels que, pour une fonction discrète quelconque sur la couche v, on ait l'inégalité

$$\| R^m v \| \leqslant \rho_m \| v \|, \qquad (119)$$

où ρ_m sont des nombres qui ne peuvent être diminués sans infraction à (119) ne serait-ce que pour une fonction v (ρ_m est appelé *norme de l'opérateur* R^m).

Estimons la grandeur u^N en utilisant (118), (119)

$$\|u^N\| \leqslant \tau \sum_{m=0}^{N-1} \|R^m f^{N-m-1}\| + \|R^N v\| \leqslant$$

$$\leqslant \tau \sum_{m=0}^{N-1} \rho_m \|f^{N-m-1}\| + \rho_N \|v\| \leqslant$$

$$\leqslant \tau N \max_m \rho_m \max_m \|f^m\| + \rho_N \|v\|. \qquad (120)$$

Comme seules nous intéressent les valeurs finies de t, on a

$$0 \leqslant m\tau < N\tau \leqslant t \quad \text{et} \quad 0 \leqslant m < N \leqslant t/\tau.$$

Il y aura stabilité si les coefficients auprès de $\max \|f^m\|$ et $\|v\|$ dans (120) resterons bornés pour $\tau, h \to 0$. Comme $\tau N \leqslant t$, la condition de stabilité s'écrira

$$\rho_m \leqslant \text{const} \quad \text{pour } \tau, h \to 0, \ m\tau \leqslant t. \qquad (121)$$

Il est évident que const ne doit pas dépendre de τ, h bien que ρ_m tout comme R et R^m dépendent de ces paramètres. C'est là l'essence même du problème car pour tous τ, h, m donnés la grandeur ρ_m est naturellement finie.

Ainsi, pour l'opérateur l de la structure stratifiée (116) la question de la *stabilité* du problème aux différences revient à l'estimation des *normes des puissances de l'opérateur* R, c'est-à-dire à la vérification de la condition (121).

Le problème envisagé au § 4 donne, de toute évidence, un exemple de l'opérateur l de la structure stratifiée. Dans ce problème nous avons d'un côté établi par estimation directe que (121) est vérifié sous la condition $-1 < a\tau/h \leqslant 0$ et d'un autre côté, nous avons fait apparaître l'instabilité à l'aide de la solution particulière $u_k \sim (-1)^k$. Dans les cas plus compliqués il est rare que l'on arrive à vérifier (121) alors que l'on peut faire apparaître l'instabilité par généralisation à un grand nombre de problèmes.

Pour étudier le comportement de $R^m v$ en fonction de m, le plus simple est d'étudier les *fonctions propres* de l'opérateur R, c'est-à-dire des fonctions discrètes sur la couche v qui, lorsqu'on leur applique R, se trouvent multipliées par des nombres λ (appelés *valeurs propres*)

$$Rv = \lambda v. \tag{122}$$

Pour les fonctions propres $R^m v = \lambda^m v$, lorsque leur nombre est suffisamment grand, la grandeur λ^m permet de caractériser la norme ρ_m.

Considérons les fonctions discrètes sur la couche $v = \{v_k\}$ déterminées et bornées pour toutes les valeurs de l'argument discret k, $-\infty < k < \infty$. Supposons que l'opérateur linéaire R soit sur ces fonctions donné par la formule

$$(Rv)_k = \sum_p \alpha_p v_{k+p}, \qquad k = 0, \pm 1, \pm 2, \ldots, \tag{123}$$

où α_p sont des coefficients donnés dépendant des paramètres τ, h; p prend un certain ensemble de valeurs. En particulier, l'exemple envisagé au § 4 est obtenu pour $\alpha_0 = 1 + \frac{a\tau}{h}$, $\alpha_1 = \frac{-a\tau}{h}$, pour les autres p on a $\alpha_p = 0$.

Les fonctions propres de l'opérateur R du type (123) doivent satisfaire à la condition (122), c'est-à-dire

$$\sum_p \alpha_p v_{k+p} = \lambda v_k, \qquad k = 0, \pm 1, \pm 2, \ldots \tag{124}$$

Nous allons chercher la solution de cette *équation* linéaire aux différences sous la forme suivante

$$v_k = v_0 q^k, \tag{125}$$

où q est un certain nombre et v_0 un facteur de normalisation. En substituant (125) dans (124) on obtient après simplification par $v_0 q^k$

$$\sum_p |\alpha_p q^p = \lambda, \tag{126}$$

c'est-à-dire que pour q quelconque la fonction v (125) satisfait à (122), avec $\lambda = \lambda(q)$ (126). De ce grand nombre de fonctions nous choisirons seulement les fonctions discrètes (125) bornées (par rapport à k). Si $|q| \neq 1$, on a $|v_k| \to \infty$ pour $k \to \infty$ ou pour $k \to -\infty$. Par conséquent $|q| = 1$.

Quant à l'équation aux différences proprement dite nous l'envisagerons seulement dans le domaine réel. Cependant toute fonction réelle peut s'écrire comme une combinaison de fonctions complexes. C'est pourquoi l'utilisation de ces dernières peut donner une information utile sur les propriétés métriques de l'opérateur R dans l'espace réel, en particulier, sur sa norme. Dans le cas présent nous avons en tout deux fonctions propres, pour $q = 1$ et $q = -1$. Le nombre de fonctions propres complexes est bien plus important et les propriétés de l'opérateur apparaissent sur ces fonctions. Posant $q = e^{i\varphi}$ on peut écrire (125), (126) comme suit

$$v_k = v_0 e^{ik\varphi}, \tag{127}$$

$$\lambda = \sum_p \alpha_p e^{ip\varphi}. \tag{128}$$

Ainsi, *à chaque* φ sur l'intervalle $(0, 2\pi)$ correspond une fonction propre v_k (127) de valeur propre λ (128). Il est évident que la grandeur v_0 ne joue aucun rôle, on peut donc poser $v_0 = 1$.

Comme pour les fonctions propres $R^m v = \lambda^m v$, on a

$$\| R^m v \| = \| \lambda^m v \| \leqslant \max_\varphi |\lambda|^m \| v \|, \tag{129}$$

la grandeur $\max |\lambda|^m$ est pour ainsi dire la norme de l'opérateur R^m sur le système de fonctions propres. Ce système est une partie de tout l'ensemble de fonctions discrètes, c'est pourquoi la norme de l'opérateur R^m ne peut être que supérieure à $\max |\lambda|^m$,

$$\max_\varphi |\lambda|^m \leqslant \rho_m. \tag{130}$$

En comparant cette inégalité avec la condition de stabilité (121), nous arrivons à la conclusion que pour vérifier cette dernière il faut que

$$\max_{\varphi} |\lambda|^m \leqslant \text{const} \quad \text{pour} \quad \tau,\ h \to 0,\ m\tau \leqslant t. \qquad (131)$$

Les valeurs propres λ, tout comme α_p en fonction desquelles elles s'expriment par (128), dépendent des paramètres τ, h. Comme $m \sim 1/\tau \to \infty$, la condition (131) équivaut à

$$\max_{\varphi} |\lambda| \leqslant 1 + O(\tau) \quad \text{pour} \quad \tau,\ h \to 0. \qquad (132)$$

Dans le cas contraire $|\lambda|^m \to \infty$. Par contre, si $|\lambda| = 1 + c\tau$, on a

$$|\lambda|^m \sim (1 + c\tau)^{t/\tau} \sim e^{ct}.$$

Cette dernière grandeur, bien qu'elle soit finie, peut être assez grande. C'est pourquoi au lieu de (132) on envisage parfois une condition plus forte

$$\max_{\varphi} |\lambda| \leqslant 1 \qquad \text{pour} \quad \tau,\ h \to 0. \qquad (133)$$

Nous avons ainsi obtenu la condition nécessaire de stabilité du problème (117) avec l'opérateur R du type (123). L'ensemble des valeurs propres λ est appelé spectre de l'opérateur et la grandeur $\max |\lambda|$, rayon spectral. Ainsi la condition (132) ou (133) est appelée *critère spectral de stabilité*.

Pour l'exemple du § 4 l'égalité (128) donne

$$\lambda = 1 + a\frac{\tau}{h} - a\,\frac{\tau}{h}\,e^{i\varphi},$$

c'est-à-dire le spectre est un cercle (dans le plan complexe λ) de rayon $|a\tau/h|$ et de centre au point $1 + a\tau/h$. Il est facile de se convaincre que $|\lambda| \leqslant 1$ seulement dans le cas où les

mêmes inégalités

$$-1 \leqslant a\frac{\tau}{h} \leqslant 0$$

sont vérifiées.

L'avantage du critère spectral de stabilité est qu'il peut facilement être généralisé à un grand nombre de problèmes plus compliqués, en particulier aux systèmes d'équations. Revenons au problème (116), (117) et supposons que les fonctions discrètes u^n, f^n, v soient des fonctions vectorielles, c'est-à-dire qu'elles soient déterminées en chaque point de calcul par plusieurs grandeurs. Tous les raisonnements ci-dessus restent vrais et seule la conclusion doit être changée. A savoir, (123) étant maintenant une égalité vectorielle, α_p sont des matrices carrées du même ordre que les vecteurs v_k. En substituant

$$v_k = v_0 q^k = v_0 e^{ik\varphi}, \tag{134}$$

où v_0 est un vecteur dans (124), on obtient

$$\sum_p \alpha_p e^{ip\varphi} v_0 = \lambda v_0$$

qui est un système d'équations linéaires homogènes par rapport aux composantes du vecteur v_0. Pour que la solurion existe, il faut que le déterminant du système soit nul:

$$\left| \sum_p \alpha_p e^{ip\varphi} - \lambda I \right| = 0, \tag{135}$$

où I est une matrice unité. La différence d'avec le cas scalaire est que maintenant le spectre se compose de plusieurs branches λ_1, λ_2, . . ., définies comme les racines de l'équation (135) au lieu de (128).

Ainsi le critère spectral permet de réduire l'étude de la stabilité au problème de l'estimation de la grandeur des racines d'une équation algébrique.

Cette équation donne dans le cas général seulement la condition nécessaire de stabilité. Cependant si pour une

certaine classe de fonctions, pour une certaine norme, le système de fonctions propres est *complet*, c'est-à-dire que toute fonction de cette classe peut être approximée par une combinaison de fonctions propres, le critère spectral est alors la condition suffisante de stabilité. Considérons par exemple le cas scalaire où l'opérateur R est du type (123), de plus nous nous bornerons aux fonctions discrètes $u = \{u_k\}$ pour lesquelles la série

$$\sum_k u_k e^{-ik\varphi} = w(\varphi) \tag{136}$$

est convergente. Chaque fonction discrète u_k donne ainsi naissance à une fonction périodique $w(\varphi)$ pour laquelle la série (136) est une série de Fourier, u_k étant les coefficients de cette série. Ces derniers, comme on sait, sont donnés par les relations

$$u_k = \frac{1}{2\pi} \int_0^{2\pi} w(\varphi)\, e^{ik\varphi}\, d\varphi \tag{137}$$

et

$$\sum_k u_k^2 = \frac{1}{2\pi} \int_0^{2\pi} |w(\varphi)|^2\, d\varphi. \tag{138}$$

On peut directement se rendre compte de la validité des relations (137), (138). A cet effet il y a lieu de multiplier (136) par $e^{im\varphi}$ dans le premier cas et par $\overline{w}(\varphi) = \sum_m u_m e^{im\varphi}$ dans le second, et d'effectuer ensuite l'intégration.

Appliquons à (137) l'opérateur R de (123), on obtient

$$(Ru)_k = \sum_p \alpha_p \frac{1}{2\pi} \int_0^{2\pi} w(\varphi)\, e^{i(k+p)\varphi}\, d\varphi =$$

$$= \frac{1}{2\pi} \int_0^{2\pi} w(\varphi)\, e^{ik\varphi} \sum_p \alpha_p e^{ip\varphi}\, d\varphi = \frac{1}{2\pi} \int_0^{2\pi} w(\varphi)\, \lambda(\varphi)\, e^{ik\varphi}\, d\varphi,$$

ceci en vertu de (128). En comparant cette égalité avec (137), on voit que $(Ru)_k$ sont les coefficients du développement en série de Fourier de la fonction $w(\varphi)\,\lambda(\varphi)$. Par conséquent pour ces coefficients on a des relations du type (138), c'est-à-dire

$$\sum_k (Ru)_k^2 = \frac{1}{2\pi}\int_0^{2\pi} |\,w(\varphi)\,\lambda(\varphi)\,|^2\,d\varphi.$$

En mettant $\max\limits_{\varphi} |\,\lambda(\varphi)\,|^2$ en facteur devant le signe de l'intégrale et en remplaçant l'intégrale restante par $\sum\limits_k u_k^2$ on obtient

$$\sum_k (Ru)_k^2 \leqslant \max_{\varphi} |\,\lambda(\varphi)\,|^2 \sum_k u_k^2. \qquad (139)$$

Définissons la norme de la fonction discrète par l'égalité

$$\|\,u\,\| = \Big(\sum_k u_k^2 h\Big)^{1/2}, \qquad (140)$$

où le facteur h a été introduit pour que, à la limite, pour $h \to 0$, la norme conserve son sens. En utilisant (140), on peut écrire (139) sous la forme

$$\|\,Ru\,\| \leqslant \max_{\varphi} |\,\lambda\,|\;\|\,u\,\|. \qquad (141)$$

Cette dernière inégalité signifie que pour la classe mentionnée de fonctions avec la norme (140), la grandeur $\max |\,\lambda\,|$ n'est pas inférieure à la norme de l'opérateur R et par conséquent le critère spectral est un critère suffisant de stabilité.

Arrêtons-nous maintenant sur une question importante pour les applications pratiques du critère spectral de stabilité. Dans tout problème réel de calcul le nombre de points de calcul est fini et l'indice k prend un ensemble fini de valeurs. Dans les points limites l'opérateur R ne peut conserver sa forme standard (123), ne serait-ce que vu l'ab-

sence d'un ensemble complet de grandeurs u_{k+p}. En ces points l'opérateur R s'exprime à l'aide de formules spéciales, réalisant telle ou telle condition aux limites. Pour pouvoir se faire une idée à quel point cette déformation peut changer les propriétés de l'opérateur R (123) standard, « sans limites », il y a lieu de revenir à l'exemple du § 4.

Supposons que R soit donné sur des fonctions discrètes u_k ($k = 0, 1, 2, \ldots, K$) par les formules

$$\left.\begin{aligned}(Ru)_k &= \left(1 + a\frac{\tau}{h}\right) u_k - a\frac{\tau}{h} u_{k+1}, \quad k = 0, 1, \ldots, K-1,\\ (Ru)_K &= 0.\end{aligned}\right\} \tag{142}$$

Il est évident que ceci correspond à un problème différentiel sur un intervalle fini $0 \leqslant x \leqslant X = Kh$ avec une condition aux limites nulle sur l'extrémité droite, dont la solution pour $a < 0$ existe et est unique.

Nous allons trouver les fonctions propres de l'opérateur R (142). Posant $Rv = \lambda v$ on arrive à un système d'équations linéaires homogènes par rapport à $v_0, v_1, \ldots, v_k$:

$$\left(\lambda - 1 - a\frac{\tau}{h}\right) v_k + a\frac{\tau}{h} v_{k+1} = 0, \qquad k = 0, 1, \ldots, K-1,$$
$$\lambda v_K = 0.$$

Il est facile de voir que ce système a deux solutions non triviales

$$v_k = \left(\frac{1 + a\frac{\tau}{h}}{a\frac{\tau}{h}}\right)^k \qquad \text{pour } \lambda = 0$$

et

$$v_0 = 1, \; v_1 = v_2 = \ldots = v_K = 0 \quad \text{pour } \lambda = 1 + a\frac{\tau}{h}.$$

Bien que la seconde solution entraîne la condition $\left|1 + a\frac{\tau}{h}\right| \leqslant 1$, il est clair qu'un nombre aussi restreint de fonctions

propres ne permet pas de mettre en évidence les propriétés de l'opérateur.

En même temps il est peu probable que la déformation de l'opérateur en un point limite puisse notablement changer ses propriétés, apparaissant d'une manière si évidente sur les fonctions $e^{ik\varphi}$. Ces dernières ne satisfont pas à la condition $u_K = 0$, c'est pourquoi nous allons les corriger, c'est-à-dire poser

$$\left.\begin{aligned} v_k &= e^{ik\varphi}, \quad k = 0, 1, \ldots, K-1, \\ v_K &= 0 \end{aligned}\right\} \tag{143}$$

et leur appliquer l'opérateur R (142). On voit aisément que pour tous les $k < K - 1$ la fonction (143) se transformera en elle-même, se trouvant multipliée par le facteur

$$\lambda = 1 + a\frac{\tau}{h} - a\frac{\tau}{h}e^{i\varphi}, \tag{144}$$

et ce n'est qu'au voisinage du point limite que cette loi cesse d'être vraie. Appliquons de nouveau l'opérateur R (142) à la fonction obtenue, c'est-à-dire étendons la correction également au point $K - 2$. En répétant ce processus on obtient pour la fonction v_k (143)

$$(R^m v)_k = \lambda^m v_k \quad \text{pour } k = 0, 1, \ldots, K - m. \tag{145}$$

Ce qui nous intéresse c'est R^m pour $\tau, h \to 0$, c'est-à-dire pour $m \sim 1/\tau \to \infty$, $K \sim 1/h \to \infty$. Si alors h/τ reste borné, en d'autres termes si m tend vers l'infini pas plus rapidement que K, il y aura toujours un intervalle de valeurs k, où (145) est vérifié. Par conséquent l'estimation inférieure de la norme $\|R^m\|$ en fonction de $|\lambda|^m$, où λ est donné par (144) c'est-à-dire est une valeur propre de l'opérateur *non pertubé*, reste vraie.

On peut raisonner autrement, en considérant au lieu de (143) la fonction

$$v_k = \left(1 - \frac{k}{K}\right)e^{ik\varphi}. \tag{146}$$

En utilisant (142), (144) on trouve que dans ce cas

$$(Rv-\lambda v)_k=\frac{a\tau}{hK}e^{i(k+1)\varphi},\ k=0,\ 1,\ \ldots,\ K-1,$$

$$(Rv-\lambda v)_K=0.$$

Comme $hK = X = \text{const}$, les dernières égalités signifient que

$$Rv - \lambda v = O(\tau) \tag{147}$$

bien que la fonction v (146) soit continue ($v_0 = 1$). On peut dire que v (146) est une fonction « presque propre » et λ (144), une valeur « presque propre » de l'opérateur R (142). Nous arrivons de nouveau à la conclusion que les conditions aux limites doivent être considérées comme une perturbation de l'opérateur, pour laquelle la plupart de ses propriétés se conservent. Il est évident qu'en se donnant des conditions aux limites absurdes on peut abîmer l'opérateur standard, mais on ne peut pas corriger ses défauts par des conditions aux limites appropriées.

Dans le cas général l'étude de ces problèmes n'est pas simple. Cependant les raisonnements ci-dessus nous montrent que pour les problèmes sur un intervalle limité la condition nécessaire de stabilité reste l'application du critère spectral à un opérateur standard limité.

Problèmes

1. A partir de la définition de la norme de l'opérateur (119) on a

$$\|R^m u\| = \|RR^{m-1}u\| \leqslant \rho_1 \|R^{m-1}u\| \leqslant \ldots \leqslant \rho_1^m \|u\|,$$

c'est-à-dire $\rho_m \leqslant \rho_1^m$. Par conséquent pour que (121) soit vérifié il suffit que $\rho_1^m \leqslant \text{const}$, ce qui équivaut à la condition suivante pour la norme de l'opérateur R

$$\rho_1 \leqslant 1 + O(\tau), \tag{148}$$

qui est par là même la condition *suffisante* de stabilité.

Montrer que pour l'opérateur R de la forme (123) la condition (148) pour la norme $\| u \| = \max_k | u_k |$ équivaut à la condition

$$\sum_p | \alpha_p | \leqslant 1 + O(\tau). \qquad (149)$$

Pour l'opérateur

$$(Ru)_k = u_k - \frac{a\tau}{2h} (u_{k+1} - u_{k-1})$$

étudier la stabilité à l'aide de (149) et du critère spectral. Comparer les résultats obtenus. Définir le problème différentiel correspondant à cet opérateur.

2. Pour le problème

$$\frac{\partial U}{\partial t} + \frac{\partial V}{\partial x} = 0, \qquad U(0, x) = U_0(x),$$

$$\frac{\partial V}{\partial t} + \frac{\partial U}{\partial x} = 0, \qquad V(0, x) = V_0(x)$$

trouver les différents schémas aux différences et étudier leur stabilité.

3. Elaborer le schéma aux différences pour la solution du problème

$$\frac{\partial U}{\partial t} + \frac{\partial f(U)}{\partial x} = \frac{\partial}{\partial x} \mu(U) \frac{\partial U}{\partial x}, \qquad U(0, x) = U_0(x),$$

où $f(U)$, $\mu(U)$ sont des fonctions données. Etudier l'approximation et la stabilité (sur le modèle linéaire).

§ 7. ÉTABLISSEMENT DES FORMULES DE CALCUL

Maintenant que nous connaissons les conditions auxquelles doit satisfaire le problème aux différences nous allons passer à son établissement. Dans chaque exemple concret nous avons simplement remplacé chaque dérivée entrant

dans une équation différentielle initiale par la relation aux différences correspondante. Mais ce n'est pas là la méthode la plus courte, ni la plus simple.

La composition du problème aux différences commence par le choix du *réseau de calcul*, c'est-à-dire d'un ensemble discret de points remplaçant le domaine continu des variations des variables indépendantes.

En principe cet ensemble est arbitraire. On peut augmenter la concentration des points des parties les plus importantes afin d'y obtenir une précision plus grande. On peut se donner la loi de formation du réseau en fonction de la solution obtenue durant le calcul. Mais sauf prescription spéciale, il est préférable de prendre le réseau régulier, déterminé par un nombre minimal de paramètres. Ceci facilite notablement l'étude du problème aux différences.

Supposons que le réseau de calcul ou la loi de sa formation soient donnés. Si l'irrégularité de réseau est peu accusée, c'est-à-dire ses paramètres varient peu d'un point à l'autre, sur de petits domaines il peut être simulé d'une manière uniforme. Ultérieurement nous considérerons un réseau régulier (pas obligatoirement rectangulaire) déterminé pour les problèmes à deux variables indépendantes x, t par deux paramètres seulement, à savoir les pas h, τ du réseau. Sur ce réseau on détermine les fonctions (ou les vecteurs fonctions) u_k^n, f_k^n, φ_k, . . . qui sont les fonctions discrètes des arguments discrets k, n, (numéros des points).

L'étape suivante est l'établissement des *équations aux différences*, c'est-à-dire des relations arithmétiques entre les grandeurs τ, h, u_k^n, f_k^n, φ_k, . . . Comme nous nous basons sur le principe de la convergence de la solution du problème discret vers la solution de l'équation différentielle, les équations aux différences doivent satisfaire à certaines conditions. Ci-dessus nous les avons formulées comme des conditions d'approximation et de stabilité.

En utilisant les formules de la dérivation numérique en vue de remplacer les dérivées par des différences finies, il

est facile d'écrire telles ou telles relations qui à la limite, pour une fonction lisse quelconque, deviendront les équations différentielles initiales. Mais ici le succès n'est pas immédiat car la majorité des schémas logiquement possibles sont instables.

La recherche d'un schéma aux différences convenable peut être plus efficace si l'on utilise un artifice que nous exposerons tout d'abord sur le même exemple simple

$$\left.\begin{aligned} \frac{\partial U}{\partial t}+a\frac{\partial U}{\partial x}=0, \\ U(x,\ 0)=U_0(x). \end{aligned}\right\} \tag{150}$$

Nous allons fixer la forme de la *maille de calcul*, c'est-à-dire indiquer les points où les valeurs de la fonction discrète sont celles que nous voulons lier aux relations aux différences. Pour le problème étudié prenons sur le réseau $x_k = kh$, $t^n = n\tau$ quatre points dont les numéros sont $(k, n+1)$, $(k-1, n)$, (k, n), $(k+1, n)$ formant une telle maille (fig. 7).

Le problème initial (150) est linéaire. Il est donc tout naturel de chercher les équations aux différences sous la forme de relations linéaires. La forme générale d'une telle relation entre les valeurs de la fonction discrète aux points mentionnés est

$$u_k^{n+1}=\alpha_{-1}u_{k-1}^n+\alpha_0u_k^n+\alpha_1u_{k+1}^n=\sum_{p=-1}^{1}\alpha_pu_{k+p}^n. \tag{151}$$

Posant $u_k^0 = U_0(x_k)$ on obtient pour tout ensemble α_p un certain problème aux différences. Nous allons choisir les coefficients α_p tels que le problème (151) donne une approximation du problème initial et soit stable.

Si pour vérifier la stabilité on a l'intention d'utiliser le critère spectral, il y a lieu d'écrire le problème (151)

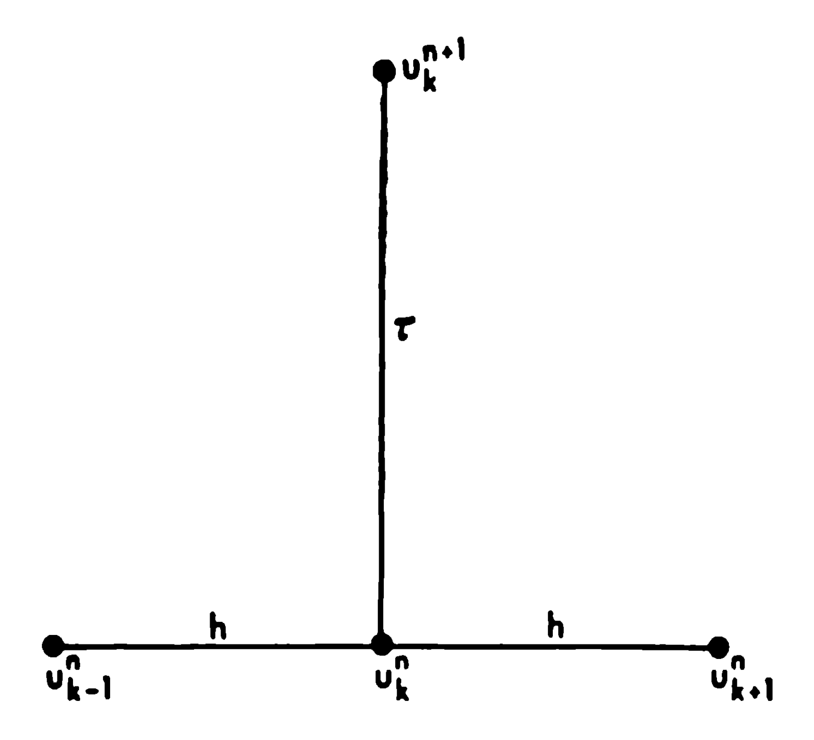


Fig. 7

sous la forme usuelle $lu = f$ en posant

$$lu = \begin{cases} \dfrac{u_k^{n+1} - \sum\limits_p \alpha_p u_{k+p}^n}{\tau}, \\ u_k^0, \end{cases} \qquad f = \begin{cases} 0, \\ U_0(x_k). \end{cases} \tag{152}$$

On sait que la condition nécessaire de stabilité s'écrit dans le cas présent sous la forme de l'inégalité suivante

$$\Big|\sum_p \alpha_p e^{ip\varphi}\Big| \leqslant 1. \tag{153}$$

Nous avons ainsi une condition imposée aux coefficients α_p.

Passons à l'approximation signifiant que $lU - f \to 0$ pour τ, $h \to 0$. Pour trouver l'ordre de grandeur de $lU - f$, utilisons comme toujours le développement de la solution de (150), à savoir de la fonction $U(x, t)$, en série suivant les puissances de τ, h, au voisinage du point central de la

cellule de calcul x_h, t^n. On a

$$\left.\begin{aligned} U_k^{n+1} &= U + \tau U_t + \frac{\tau^2}{2} U_{tt} + \frac{\tau^3}{6} U_{ttt} + \ldots, \\ U_{k+p}^n &= U + phU_x + \frac{p^2h^2}{2} U_{xx} + \frac{p^3h^3}{6} U_{xxx} + \ldots, \end{aligned}\right\} \qquad (154)$$

où U, U_t, . . . sont pris au point central. Comme $U(x, t)$ est la solution de l'équation (150) pour cette fonction on a les égalités

$$\left.\begin{aligned} U_t &= -aU_x, \\ U_{tt} &= a^2U_{xx}, \\ U_{ttt} &= -a^3U_{xxx}, \\ &\ldots\ldots\ldots \end{aligned}\right\} \qquad (155)$$

obtenues par dérivation de (150).

En utilisant (154), (155) on peut trouver la combinaison qui nous intéresse des valeurs de U:

$$U_k^{n+1} - \sum_p \alpha_p U_{k+p}^n = \left(1 + \sum \alpha_p\right) U - \left(\tau a + h \sum p\alpha_p\right) U_x + \\ + \frac{1}{2}\left(\tau^2a^2 - h^2 \sum p^2\alpha_p\right) U_{xx} - \\ - \frac{1}{6}\left(\tau^3a^3 + h^3 \sum p^3\alpha_p\right) U_{xxx} + \ldots \qquad (156)$$

ne différant de $lU - f$ que par le facteur $1/\tau$. L'ordre de l'approximation dépendra des premiers termes non nuls de ce développement. Les grandeurs U, U_x, U_{xx}, . . . doivent être supposées quelconques et indépendantes. En annulant les coefficients de ces grandeurs, on obtient une chaîne

d'équations

$$\left.\begin{aligned}1-\sum\alpha_p=0,\\ \tau a+h\sum p\alpha_p=0,\\ \tau^2a^2-h^2\sum p^2\alpha_p=0,\\ \tau^3a^3+h^3\sum p^3\alpha_p=0,\\ \ldots\ldots\ldots\ldots\end{aligned}\right\}\qquad(157)$$

qui sont les équations permettant de déterminer α_p dont l'ordre, par rapport à τ, h, augmente. Plus grand est le nombre d'équations auquel on peut satisfaire, plus l'ordre de l'approximation sera élevé.

Nous n'avons à notre disposition que trois coefficients non déterminés α_p. C'est pourquoi nous pouvons satisfaire à trois des équations (157) au plus. En désignant $a\tau/h$ par r, on peut écrire ces équations comme suit

$$\alpha_{-1}+\alpha_0+\alpha_1=1,\quad \alpha_{-1}-\alpha_1=r,\quad \alpha_{-1}+\alpha_1=r^2.$$

En les résolvant, on obtient

$$\alpha_{-1}=\frac{r^2+r}{2},\ \alpha_0=1-r^2,\ \alpha_1=\frac{r^2-r}{2}.\qquad(158)$$

Pour ces valeurs de α_p l'ordre de grandeur du premier terme différent de zéro dans le développement (156) est τ^3, rh^3. Par conséquent $lU-f=0\,(\tau^2,\,h^2)$, c'est-à-dire l'approximation est du second ordre par rapport à τ, h.

Il ne reste qu'à satisfaire la condition de stabilité (153). En substituant (158) dans (153), on obtient :

$$\begin{aligned}\sum\alpha_pe^{ip\varphi}&=\frac{r^2+r}{2}e^{-i\varphi}+1-r^2+\frac{r^2-r}{2}e^{i\varphi}=\\ &=1-r^2+r^2\cos\varphi-ir\sin\varphi=\\ &=1-2r^2\sin^2\frac{\varphi}{2}-i\,2r\sin\frac{\varphi}{2}\cos\frac{\varphi}{2}.\end{aligned}$$

Par conséquent

$$\left|\sum \alpha_p e^{ip\varphi}\right|^2 = \left(1 - 2r^2 \sin^2 \frac{\varphi}{2}\right)^2 + \left(2r \sin \frac{\varphi}{2} \cos \frac{\varphi}{2}\right)^2 =$$

$$= 1 - 4r^2 \sin^2 \frac{\varphi}{2} + 4r^4 \sin^4 \frac{\varphi}{2} + 4r^2 \sin^2 \frac{\varphi}{2} \cos^2 \frac{\varphi}{2} =$$

$$= 1 - 4r^2 \sin^4 \frac{\varphi}{2} + 4r^4 \sin^4 \frac{\varphi}{2} =$$

$$= 1 - 4r^2 (1 - r^2) \sin^4 \frac{\varphi}{2}.$$

Il est évident que pour vérifier l'inégalité (153) il faut que $r^2 \leqslant 1$, c'est-à-dire

$$|a| \frac{\tau}{h} \leqslant 1. \tag{159}$$

C'est la condition de stabilité du problème aux différences (151), (158).

Si pour trouver les coefficients α_p on se borne aux deux premières équations (157), le schéma obtenu sera de toute évidence une approximation du premier ordre $O(\tau, h)$. Comme ces deux équations contiennent trois inconnues α_p, il y a un grand nombre de schémas pouvant être utilisés. L'un de ces schémas a été envisagé au § 4. Si l'on utilise seulement la première des équations (157), l'approximation ne se trouve pas assurée, car dans ce cas l'erreur sera finie.

Il va de soi que la méthode décrite peut également être utilisée pour trouver les formules de calcul d'autres problèmes. Avant tout, un raisonnement à priori nous permet de choisir la forme de la maille de calcul et celle des relations aux différences contenant des *coefficients indéterminés.* En exigeant ensuite que soient remplies les conditions d'approximation et de stabilité, le problème revient à la solution d'un système d'équations et d'inégalités algébriques.

Lorsque l'on utilise la méthode des coefficients indéterminés pour la solution des problèmes compliqués on est obligé de faire des calculs analytiques compliqués.

Pour simplifier les calculs il y a lieu de renoncer à la généralité et de se borner à la solution des formes simples d'équations aux différences. Par exemple dans le problème envisagé, on aurait pu au lieu de (151) chercher le schéma aux différences sous la forme

$$\frac{u_k^{n+1}-u_k^n}{\tau}=c_1\frac{u_{k+1}^n-u_k^n}{h}+c_2\frac{u_k^n-u_{k-1}^n}{h}, \qquad (160)$$

les coefficients c_1, c_2 étant indéterminés.

Il est évident que si les équations aux différences satisfaisant aux conditions imposées (forme de la maille de calcul, ordre de l'approximation, forme de la relation aux différences) existent, elles peuvent toutes être obtenues par la manière décrite.

Nous allons aborder encore une question. Lors de l'étude du schéma aux différences (151) nous l'avons écrit sous la forme $lu = f$, de plus, nous avons déterminé l par la formule (152) et pour une raison « mystérieuse » divisé (151) par τ. Il semble plus naturel de définir l par l'égalité suivante

$$lu=u_k^{n+1}-\sum_{p=-1}^{1}\alpha_p u_{k+p}^n. \qquad (161)$$

Remarquons que dans ce cas (pour les mêmes α_p) on obtiendra $lU - f = \tau O(\tau^2, h^2)$ c'est-à-dire une approximation d'ordre plus élevé. Cependant les propriétés du problème ne dépendent pas, de toute évidence, des désignations, des notations et de la méthode d'étude. L'avantage d'une meilleure approximation est illusoire et se perd vu la stabilité plus faible. En effet, comme il est facile de le voir, pour le problème $lu = f$ où l est donné par (161) on obtient l'estimation $u \sim f/\tau$, ce qui, conformément à notre définition, signifie qu'il y a instabilité. Néanmoins ceci n'a pas d'influence sur la convergence car on a un ordre de réserve dans l'approximation.

La remarque faite montre la non-unicité formelle apparaissant lorsque l'on divise le problème de la convergence en approximation et stabilité.

Examinons encore une méthode permettant de trouver les formules de calcul. Cette méthode a un domaine d'application plus restreint mais se trouve être souvent suffisamment efficace.

On peut dire que notre méthode d'approximation des problèmes différentiels est *locale*. Lors de l'établissement des équations aux différences toute notre attention se porte sur une maille de calcul, sur le domaine voisin du point de calcul dont les dimensions sont de l'ordre de τ, h. Or, de même que toute fonction lisse peut localement être supposée linéaire, tout problème envisagé dans un petit domaine où la solution varie peu peut être approximé par un problème linéaire à coefficients constants.

Par exemple au lieu de l'équation

$$\frac{\partial U}{\partial t}+a(x,\ t,\ U)\frac{\partial U}{\partial x}=0, \tag{162}$$

on peut au voisinage du point x_k, t^n considérer l'équation

$$\frac{\partial U}{\partial t}+a_k^n\frac{\partial U}{\partial x}=0, \tag{163}$$

où $a_k^n = a\,(x_k,\ t^n,\ U_k^n)$. En effet, si l'on substitue la solution (lisse) de la première équation dans la seconde, celle-ci est satisfaite à $O\,(\tau,\ h)$ près car

$$(a\,(x,\ t,\ U\,(x,\ t))-a_k^n)\,\frac{\partial U}{\partial x}=O\,(\tau,\ h).$$

Ajoutons à ce qui vient d'être dit que l'étude de la stabilité des problèmes aux différences (§ 5) que nous avons faite ci-dessus est basée sur leurs modèles linéaires. Toute la théorie de la stabilité concerne en fait la stabilité des équations linéaires aux différences.

Les raisonnements ci-dessus montrent qu'il y a lieu d'envisager les méthodes d'établissement des formules de calcul en nous limitant aux problèmes linéaires à coefficients constants.

Ci-dessus, lorsque nous avons exposé différentes propriétés sur l'exemple du problème (150), nous avons systématiquement évité d'utiliser le fait que nous connaissons la solution exacte et que celle-ci est donnée par une formule explicite, à savoir

$$U(x, t) = U_0 (x - at). \tag{164}$$

En l'utilisant nous risquions d'obtenir des affirmations vraies seulement pour les problèmes dont on connaît la solution, notre étude n'ayant plus alors de sens. Nous avons fait appel au problème (150) qui représente une certaine classe de problèmes et nous avons utilisé celles de ses propriétés qui sont typiques pour cette classe. Par contre l'existence d'une formule donnant l'expression explicite de la solution caractérise les équations linéaires à coefficients constants.

Comme à l'heure actuelle ce sont justement des problèmes de ce genre qui nous intéressent, on est en droit d'utiliser cette propriété. Voyons la manière de procéder d'abord sur l'exemple de l'équation (163).

Nous avons besoin d'obtenir une relation exprimant u_k^{n+1} en fonction des grandeurs de la n-ième couche u_k^n, $u_{k\pm1}^n$, . . . La formule de la solution exacte (164) appliquée au cas envisagé donne

$$U(x_k, t^{n+1}) = U(x_k - a_k^n \tau, t^n). \tag{165}$$

Par conséquent, pour trouver u_k^{n+1} il faut connaître la valeur de la solution au point $x = x_k - a_k^n \tau$ de la n-ième couche (fig. 8). Mais sur cette couche nous avons un ensemble *discret* de valeurs u_k^n, $u_{k\pm1}^n$, . . ., c'est-à-dire la fonction discrète. Pour calculer la valeur dont nous avons besoin, procédons à l'interpolation.

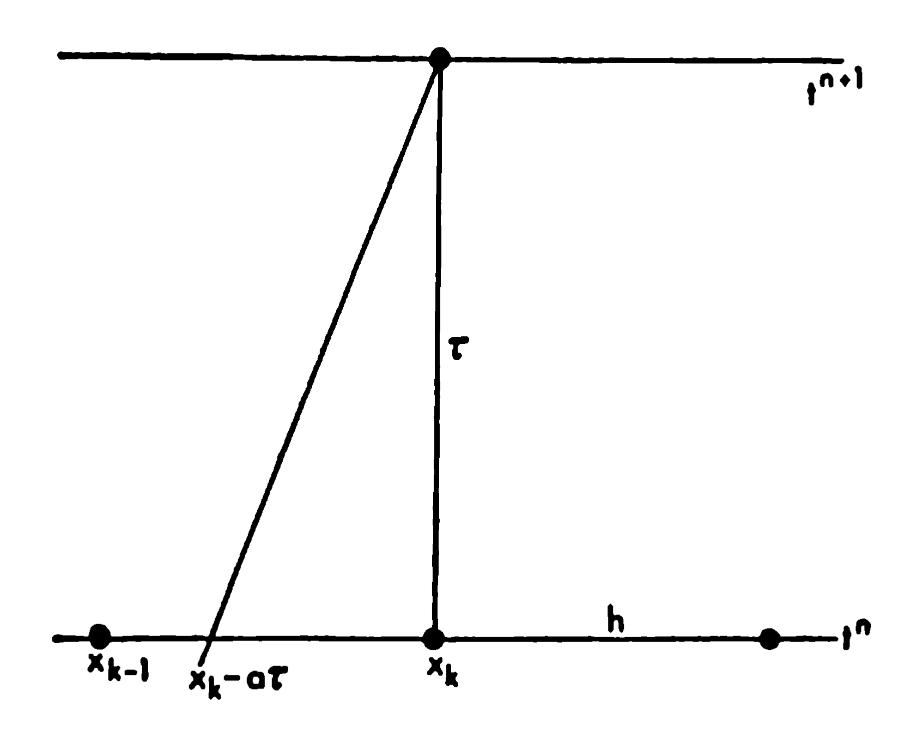


Fig. 8

Supposons que

$$x_{k-1} \leqslant x_k - a_k^n \tau \leqslant x_k. \qquad (166)$$

L'interpolation linéaire donne alors

$$u^n (x_k - a_k^n \tau) = a_k^n \frac{\tau}{h} u_{k-1}^n + \left(1 - a_k^n \frac{\tau}{h}\right) u_k^n, \qquad (167)$$

et conformément à la formule (165) on obtient

$$u_k^{n+1} = u_k^n - a_k^n \frac{\tau}{h} (u_k^n - u_{k-1}^n),$$

qui est l'équation aux différences que nous connaissons déjà et qui est stable pour

$$0 \leqslant a_k^n \frac{\tau}{h} \leqslant 1.$$

On voit ainsi que la condition de stabilité dans le cas présent coïncide avec la condition (166) assurant l'interpolabilité de la formule (167). Ce n'est pas étonnant car pour assurer la stabilité il faut prendre en considération le domaine

de dépendance de la solution. C'est là justement ce qui est essentiel pour la méthode proposée.

Maintenant que les traits essentiels de la méthode sont clairs, sans aspirer à la généralité, nous allons les exposer pour le cas où le problème initial consiste à intégrer un système d'équations différentielles aux dérivées partielles du type

$$\frac{\partial U}{\partial t} = \mathscr{D}(U) \tag{168}$$

pour des conditions initiales données. U est ici le vecteur fonction de x, t cherché et $\mathscr{D}$ un opérateur différentiel (par rapport à x) pour lequel le problème posé est correct, c'est-à-dire sa solution existe, elle dépend d'une manière unique et continue des données initiales. Ces problèmes sont dits *d'évolution* car ils décrivent le développement dans le temps d'un certain état initial.

La première étape consiste à créer un modèle linéaire du système donnant une approximation locale du système (168). L'opérateur $\mathscr{D}$ est une certaine fonction vecteur donnée des arguments U, U_x, U_{xx}, ...

$$\mathscr{D}(U) = f(U, U_x, U_{xx}, \ldots).$$

Donc au voisinage des valeurs U_k^n, $(U_x)_k^n$ $(U_{xx})_k^n$, ... on peut approximer l'opérateur $\mathscr{D}$ par l'expression linéaire suivante

$$\mathscr{D}(U) \sim f_k^n + (f_U)_k^n (U - U_k^n) + (f_{U_x})_k^n (U_x - (U_x)_k^n) + \ldots \tag{169}$$

f_U, f_{U_x}, ... désignent les matrices des dérivées des composantes du vecteur f par rapport aux composantes des vecteurs U, U_x, ... et f_k^n, U_k^n, $(U_x)_k^n$ sont les valeurs des grandeurs correspondantes au point x_k, t^n.

Supposant que la solution $U(x, t)$ soit lisse, c'est-à-dire que les différences $U - U_k^n$, $U_x - (U_x)_k^n$, ... soient petites, au lieu de (168) considérons le système d'équations différen-

tielles linéaires à coefficients constants

$$\frac{\partial U}{\partial t} = DU + C, \qquad (170)$$

où D est un opérateur différentiel linéaire, s'exprimant par la partie homogène (169), et C une constante tenant compte des termes restant de (169):

$$DU = (f_U)_k^n U + (f_{U_x})_k^n U_x + \ldots,$$

$$C = f_k^n - (f_U)_k^n U_k^n - (f_{U_x})_k^n (U_x)_k^n + \ldots$$

La théorie des équations différentielles aux dérivées partielles nous apprend que la solution d'un système d'équations linéaires à coefficients constants (170) peut s'exprimer par la formule suivante

$$U(x, t) = QU(x, 0) + Ct, \qquad (171)$$

où Q est un opérateur intégral linéaire, à savoir

$$QU(x, 0) = \int_{-\infty}^{\infty} q(\xi, t)\, U(x+\xi, 0)\, d\xi, \qquad (172)$$

et q la matrice des fonctions correspondant au système (170). Nous n'allons pas nous arrêter ici sur les méthodes permettant de l'obtenir. Mentionnons par exemple que pour l'équation de la *conductibilité thermique*

$$\frac{\partial U}{\partial t} = a\frac{\partial^2 U}{\partial x^2}, \qquad a > 0, \qquad (173)$$

la fonction q est de la forme

$$q(x, t) = \frac{1}{2\sqrt{\pi a t}} e^{-\frac{x^2}{4at}}. \qquad (174)$$

Pour résoudre le problème (150) on peut également écrire la formule (164) sous la forme (171), (172), ceci à condition de

poser

$$q(x, t) = \delta(x + at)$$

où $\delta(x)$ est la fonction delta.

La formule (171) permet d'exprimer la solution à l'instant $t^{n+1} = t^n + \tau$ à l'aide de la fonction $U(x, t^n)$. Dans notre cas cette dernière est représentée par la fonction *discrète* u_k^n. Donc, pour utiliser la formule (171) trouvons tout d'abord la fonction *d'interpolation* $Pu^n(x)$. Comme toute notre étude porte sur le voisinage du point x_k, t^n, nous écrirons cette fonction comme suit

$$Pu^n(x) = \sum_m u_{k+m}^n P_m(x), \tag{175}$$

où $P_m(x)$ sont les polynômes correspondants.

La fonction d'interpolation (175) peut également être utilisée pour la détermination des valeurs $(U_x)_k^n, \ldots$, entrant dans l'expression (169), et par conséquent, dans les coefficients D et le deuxième terme du deuxième membre C de (170).

En substituant $Pu^n(x)$ dans (172) on obtient

$$QPu^n(x) = \sum_m \alpha_m u_{k+m}^n,$$

où

$$\alpha_m = \int_{-\infty}^{\infty} q(\xi, \tau) P_m(x_k + \xi)\, d\xi,$$

et en vertu de (171) on peut écrire

$$u_k^{n+1} = \sum_m \alpha_m u_{k+m}^n + C\tau,$$

c'est-à-dire la formule de calcul dont on a besoin. Il est évident que cette dernière est simplement une formule de quadrature pour (171), (172).

Nous allons faire plusieurs remarques. Le système (170) donne une approximation du système initial (168) à des termes du second ordre par rapport à τ, h près, car l'erreur de l'approximation linéaire de (169) est du second ordre par rapport à $U - U_k^n$, $U_x - (U_x)_k^n$, . . . Afin de simplifier l'opérateur linéaire D, on peut en écrivant $\mathcal{D}$ (U) sous la forme (169), ne tenir compte que des dérivées d'ordre élevé (comme dans (162), (163)) car la stabilité dépend essentiellement de la forme des termes des équations aux différences correspondant aux dérivées d'ordre supérieur. Il est vrai que ceci a pour effet d'abaisser l'ordre de l'approximation jusqu'à O (τ, h).

Pour les mêmes raisons, lors du calcul des grandeurs U_k^n, $(U_x)_k^n$, . . ., entrant dans les coefficients et dans le second membre de (170), il n'est pas obligatoire d'utiliser la fonction d'interpolation (175). A cet effet on peut utiliser des expressions quelconques aux différences donnant évidemment une approximation de ces grandeurs.

Tous les schémas aux différences obtenus par la méthode exposée ne diffèrent les uns des autres que par la méthode d'approximation locale (169) et par la forme de l'interpolation (175). Si l'interpolation utilisée est suffisamment précise et permet de tenir compte du domaine de dépendance de la solution d'une manière efficace, le schéma aux différences est satisfaisant. Son étude s'effectue alors par les méthodes habituelles.

Problèmes

1. Quelle est l'imprécision, en tant qu'ordre par rapport à τ, h, que l'on peut admettre lors de la résolution du système (157)?

2. Trouver tous les schémas stables du type (151) et (160) pour le problème (150) du premier ordre par rapport à τ, h.

3. En utilisant la méthode des coefficients indéterminés trouver le schéma aux différences du type (151) pour la

solution de l'équation de la conductibilité thermique (173). Pour quelle relation entre τ et h ce schéma donnera-t-il une précision maximale?

4. En utilisant la formule de la solution exacte (171), (172), (174) de l'équation de la conductibilité thermique (173), construire le schéma aux différences, en utilisant pour $t = t^n$:

a) l'interpolation linéaire sur deux points x_{k-1}, x_{k+1};

b) l'interpolation linéaire par morceaux sur x_{k-1}, x_k pour $x < x_k$ et sur x_k, x_{k+1} pour $x > x_k$;

c) l'interpolation quadratique sur trois points x_{k-1}, x_k, x_{k+1}.

Etudier l'approximation et la stabilité de chacun des schémas obtenus.

5. La solution du système d'équations

$$\frac{\partial U}{\partial t}+\frac{\partial V}{\partial x}=0,$$

$$\frac{\partial V}{\partial t}+\frac{\partial U}{\partial x}=0$$

est donnée par les formules

$$U(x,\ t)=\frac{1}{2}\left(U(x+t,\ 0)+U(x-t,\ 0)-V(x+t,\ 0)+V(x-t,\ 0)\right),$$

$$V(x,\ t)=\frac{1}{2}\left(V(x+t,\ 0)+V(x-t,\ 0)-U(x+t,\ 0)+U(x-t,\ 0)\right).$$

Les utiliser pour construire le schéma aux différences correspondant au système

$$\frac{\partial U}{\partial t}+\frac{\partial p(V)}{\partial x}=0,$$

$$\frac{\partial V}{\partial t}=\frac{\partial U}{\partial x}=0,$$

où $p(V)$ est une fonction donnée.

§ 8. SCHÉMAS AUX DIFFÉRENCES IMPLICITES

Lors de la résolution d'un problème concret par une méthode numérique il y a lieu de choisir les valeurs des paramètres de la méthode, c'est-à-dire les pas du réseau τ, h, soit la quantité et la disposition des points de calcul. La solution de cette question dépend d'un grand nombre de facteurs différents : des propriétés de la solution, des exigences à la précision des calculs, des moyens disponibles de réalisation de la méthode, c'est-à-dire de la puissance des ordinateurs utilisés, etc. C'est pourquoi on ne peut donner une méthode pouvant être utilisée pour un grand nombre de problèmes.

L'étude théorique de la convergence de la méthode n'est effectuée que pour résoudre des questions de principe. Pour les problèmes réels τ et h ne peuvent tendre vers zéro. C'est pourquoi en choisissant le réseau de calcul de telle façon que la précision soit suffisante, on se base, en général, sur une certaine information sur les propriétés caractéristiques de la solution et l'on exige que le réseau soit suffisamment dense pour que ces propriétés puissent apparaître ainsi que les lois qui nous intéressent du comportement de la solution. A cet effet, en règle générale, on n'a pas besoin d'avoir un grand nombre de points.

Cependant ce raisonnement n'est pas vrai pour toutes les méthodes numériques car il ne tient pas compte des propriétés individuelles, internes de la méthode. Sans toucher à tous les aspects de cette question, nous nous arrêterons sur le plus simple.

Toutes les équations aux différences concrètes que nous avons étudiées ci-dessus se sont trouvées être stables pour certaines conditions du type

$$\frac{\tau}{h} < \text{const}, \quad \frac{\tau}{h^2} < \text{const} \tag{176}$$

imposées aux pas du réseau.

Il est évident que lorsque l'on choisit les paramètres du réseau de calcul on doit en tenir compte. Souvent pour les valeurs de τ, h donnant la précision requise les conditions (176) ne sont pas vérifiées. C'est pourquoi on est obligé de diminuer τ (et ceci notablement), ce qui augmente considérablement la quantité de calculs.

Les conditions (176) étant exprimées à l'aide des paramètres de la méthode numérique et non de ceux du problème initial et comme elles caractérisent cette méthode, on peut essayer d'élaborer des méthodes imposant des conditions plus faibles aux pas du réseau ou mêmes sans ces conditions.

Les schémas aux différences étudiés ci-dessus donnaient l'expression explicite en chaque point d'une couche temporelle donnée, u_k^{n+1}, en fonction des valeurs de la solution en plusieurs points voisins de la couche précédente u_k^n, $u_{k\pm1}^n$, c'est pourquoi ces schémas sont dits *explicites*. Pour ces schémas les conditions de stabilité du type (176) expriment qu'il est nécessaire de tenir correctement compte du domaine de dépendance de la solution. Cette relation n'est pas toujours aussi apparente que dans l'exemple du § 4.

Considérons le problème

$$\left.\begin{array}{c}\dfrac{\partial U}{\partial t}=a\,\dfrac{\partial^2 U}{\partial x^2}, \quad U(0,\ x)=U_0(x),\\ -\infty<x<\infty, \quad 0\leqslant t\leqslant T\end{array}\right\} \tag{177}$$

et le schéma aux différences donnant son approximation

$$\left.\begin{array}{c}\dfrac{u_k^{n+1}-u_k^n}{\tau}=a\,\dfrac{u_{k+1}^n-2u_k^n+u_{k-1}^n}{h^2}, \quad u_k^0=U_0(x_k),\\ k=0,\ \pm1,\ \pm2,\ \ldots,\quad n=0,\ 1,\ \ldots,\ N.\end{array}\right\} \tag{178}$$

En utilisant le critère spectral, on peut facilement voir que la condition de stabilité du problème (178) s'écrit

$$\frac{\tau}{h^2}\leqslant\frac{1}{2a}. \tag{179}$$

Comparons les domaines de dépendance de la solution des problèmes (177) et (178). On sait que (voir (172) à (174)) pour l'équation de la conductibilité thermique (177) ce domaine est la droite $-\infty < x < \infty$ toute entière. Mais dans le cas du problème aux différences (178) le domaine de dépendance pour les points de la N-ième couche seront les points de la couche zéro, remplissant l'intervalle de largeur finie $2Nh$ (fig. 9). Ainsi, bien que l'on ne tienne que partiellement compte du domaine de dépendance de la solution exacte, le problème (178) sous la condition (179) est stable. La contradiction se trouve évincée par le fait que pour $\tau, h \to 0$, sous la condition (179), la largeur de l'intervalle mentionné croît indéfiniment. En effet, si $\tau = T/N$ en vertu de (179) on a $T/Nh^2 \leqslant 1/2a$ et

$$Nh \geqslant \frac{2aT}{h} \to \infty \text{ pour } \tau,\ h \to 0.$$

Par conséquent à la limite on tient compte correctement du domaine de dépendance de la solution, ce qui de nouveau est lié aux limitations imposées par l'inégalité (179) aux pas du réseau.

Ainsi, les schémas aux différences n'ayant pas de limitations importantes sur les pas du réseau, doivent être assez compliqués. Pour tenir compte d'une manière efficace du domaine de dépendance, il faut lors du calcul des grandeurs aux points d'une couche donnée utiliser un grand nombre de points de la couche précédente. Abordons la cause de l'instabilité et de l'apparition de la condition (179) d'un autre côté, quelque peu formel.

En vérifiant la stabilité du schéma aux différences (178) à l'aide du critère spectral, nous étudions en fait le comportement des solutions particulières du type

$$u_k^n = \lambda^n e^{ik\varphi}, \tag{180}$$

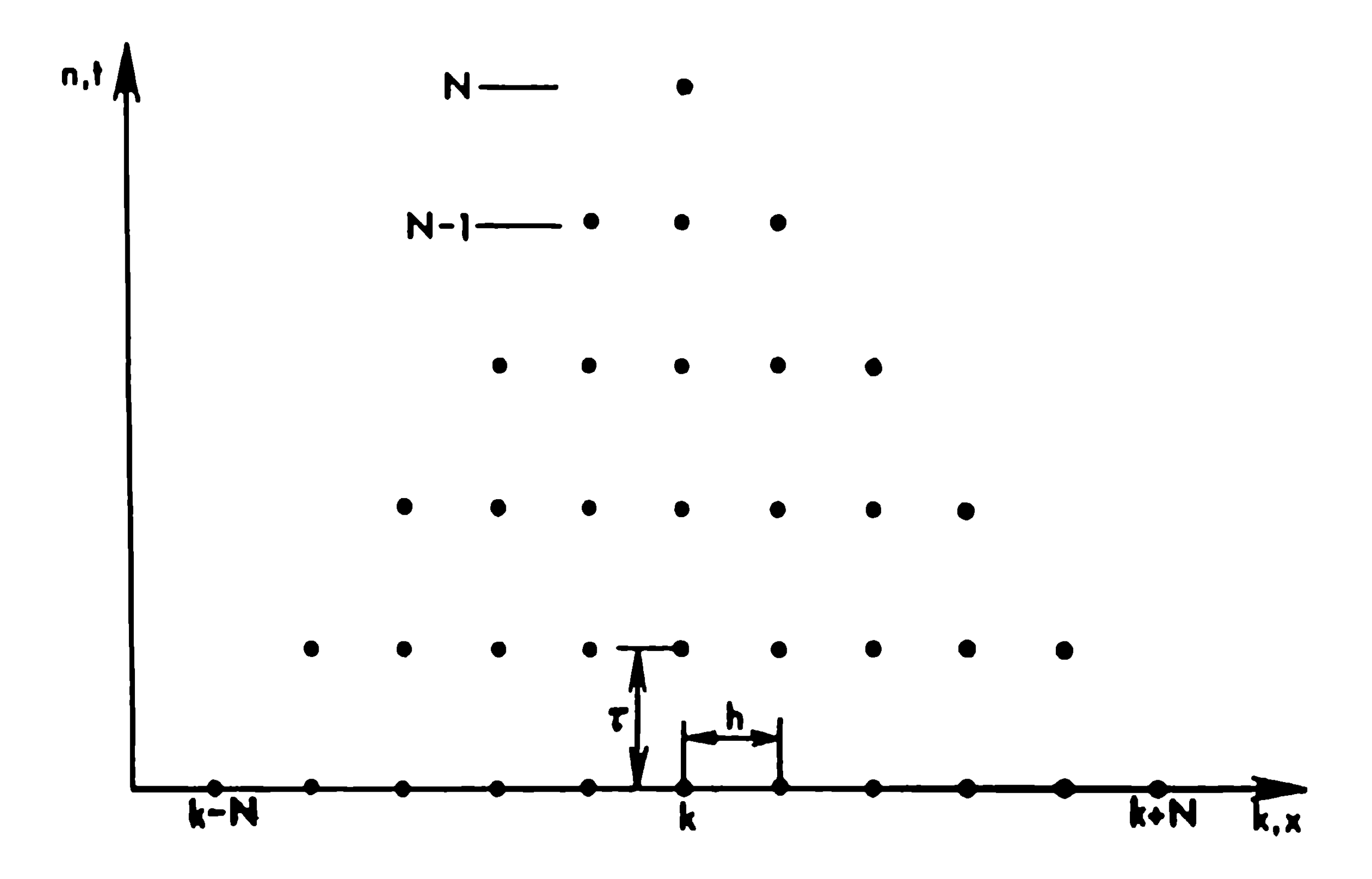


Fig. 9

c'est-à-dire utilisons la méthode de Fourier. En substituant (180) dans l'équation aux différences (178), on arrive à

$$\frac{\lambda^{n+1}-\lambda^{n}}{\tau}=-M\lambda^{n}, \tag{181}$$

où

$$M=2a\frac{1-\cos\varphi}{h^{2}}, \tag{182}$$

d'où l'on tire

$$\lambda^{n+1}=(1-\tau M)\lambda^{n}=(1-\tau M)^{n+1}, \tag{183}$$

et comme seul le cas $|\lambda|\leqslant 1$ nous intéresse, on obtient la condition $|1-\tau M|\leqslant 1$, c'est-à-dire (179).

L'exposant n dans (181) peut être interprété comme un indice (numéro du pas suivant t) et la relation (181) comme une équation aux différences correspondant à l'équation

différentielle

$$\frac{d\lambda}{dt} = -M\lambda, \qquad \lambda(0) = 1, \tag{184}$$

où M est une constante positive, aussi grande que l'on veut, vu que h est petit.

La solution exacte de l'équation (184) est

$$\lambda = e^{-Mt}, \tag{185}$$

et la solution approchée donnée par (181) c'est-à-dire (183), convergera vers elle pour $\tau \to 0$, car $(1 - \tau M)^{t/\tau} \to e^{-Mt}$. De ce point de vue la méthode (181) (c'est simplement la méthode d'Euler) d'intégration de l'équation (184) est acceptable. Mais cette méthode a un autre inconvénient. Alors que pour la solution exacte (185) l'estimation $|\lambda| \leqslant 1$ est vraie pour $M > 0$ quelconque, pour la solution approchée (183) elle ne l'est que pour des M petits satisfaisant à l'inégalité $|1 - \tau M| \leqslant 1$. Mais c'est justement cette propriété de la solution qui nous intéresse.

On peut facilement comprendre la cause de cet inconvénient : la solution exacte (185) est une fonction à décroissance rapide (pour M grand) et la méthode (181) utilise pour le calcul de la dérivée sa valeur disparaissant instantanément, non utilisable pour le pas suivant ($-M\lambda^n$ dans le second membre de (181)) (fig. 10). On peut rectifier la situation en prenant la valeur de la dérivée d'un pas à l'avance $-M\lambda^{n+1}$, c'est-à-dire

$$\frac{\lambda^{n+1} - \lambda^n}{\tau} = -M\lambda^{n+1}. \tag{186}$$

Dans ce cas au lieu de (183) on obtient

$$\lambda^{n+1} = \frac{\lambda^n}{1 + \tau M} = \frac{1}{(1 + \tau M)^{n+1}}. \tag{187}$$

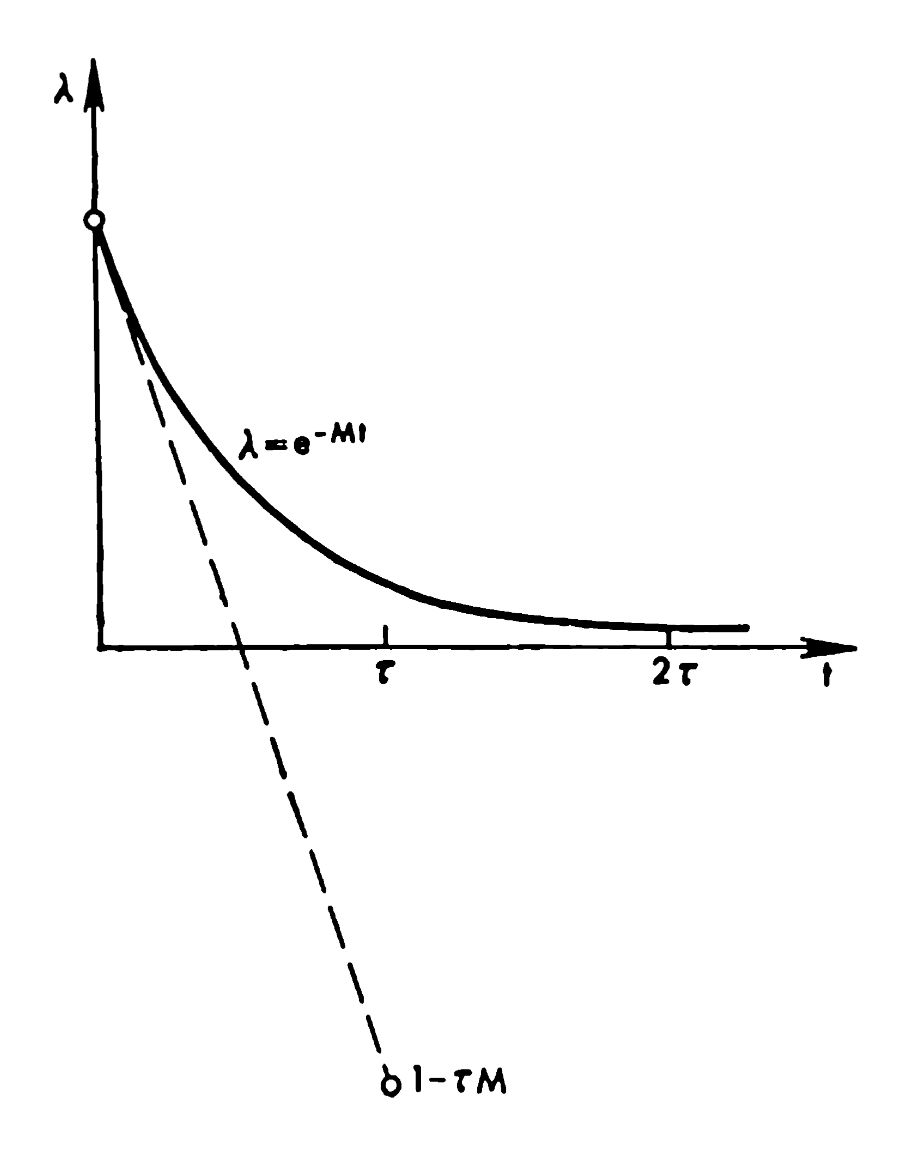


Fig. 10

Cette solution, tout comme (183), converge de toute évidence vers la solution exacte pour $\tau \to 0$, mais à la différence de (183) elle satisfait à la condition $|\lambda^{n+1}| \leqslant 1$ pour tout $M > 0$ aussi grand que l'on veut, c'est-à-dire laisse bien voir la propriété essentielle de la solution exacte, à savoir sa décroissance rapide.

Revenant au problème (177), on peut dire que la condition de stabilité (179) est levée si on élabore un schéma aux différences correspondant à (186) et non à (181). En tout cas, la vérification de ce schéma sur des fonctions du type (180) ne fait pas ressortir d'instabilité. On a de toute évidence le

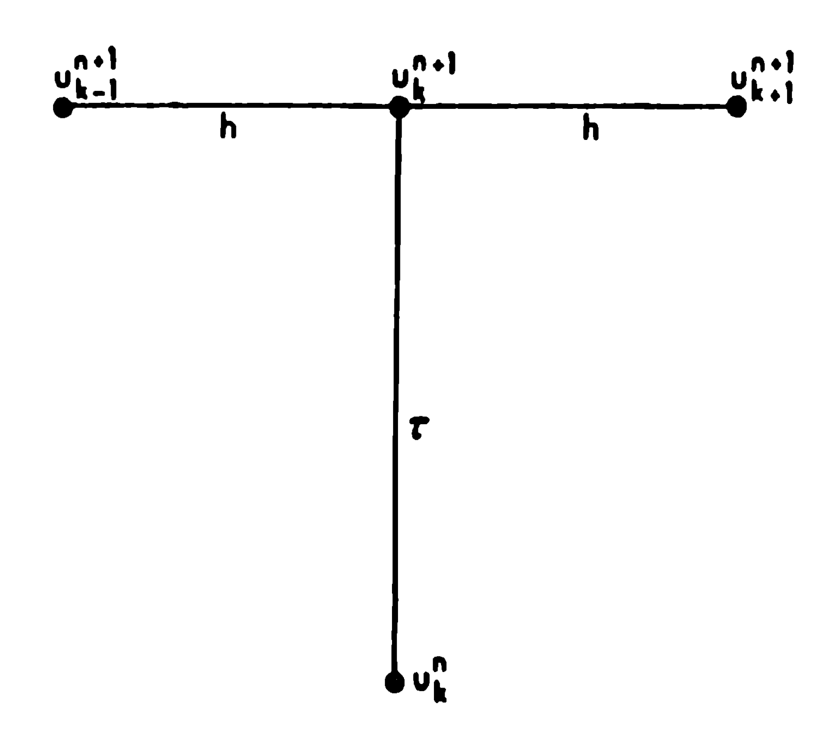


Fig. 11

schéma suivant :

$$\left.\begin{array}{c}\dfrac{u_k^{n+1}-u_k^n}{\tau}=a\,\dfrac{u_{k+1}^{n+1}-2u_k^{n+1}+u_{k-1}^{n+1}}{h^2},\\ u_k^0=U_0(x_k),\\ k=0,\ \pm1,\ \pm2,\ \ldots,\quad n=0,\ 1,\ \ldots,\ N,\end{array}\right\}\qquad(188)$$

qui ne diffère de (178) qu'en ce que les grandeurs dans le second membre sont prises non pas sur la n-ième, mais sur la $(n+1)$-ième couche ; la maille de calcul correspondante est donnée fig. 11. Il en découle que chacune des équations (188) reliant les valeurs u_k^{n+1} en trois points ne donne pas une expression explicite pour u_k^{n+1}, pour trouver ces dernières il faut résoudre un système d'un nombre peut-être pas infini, mais en tout cas grand, d'équations linéaires (188). Ces schémas sont dits *implicites*. Nous exposerons plus loin les méthodes de résolution du système (188), maintenant nous allons continuer son étude.

Il est évident que le problème (188), tout comme (178), donne une approximation du problème différentiel initial

(177). La vérification de la stabilité à l'aide de la fonction (180) est basée sur le critère spectral. Nous l'avons formulé seulement pour des schémas explicites, son utilisation pour des schémas implicites doit être argumentée. Cependant dans le cas du problème (188) on peut démontrer la stabilité directement.

Modifions quelque peu le problème pour en faire le calcul moins ardu. Ainsi, nous allons considérer les équations aux différences (188) seulement pour un ensemble fini (même très grand) de valeurs k et les compléter aux points extrêmes par des conditions aux limites en donnant par exemple les valeurs de la fonction en ces points. On obtient le système

$$\left.\begin{aligned} \frac{u_k^{n+1}-u_k^n}{\tau}-a\,\frac{u_{k+1}^{n+1}-2u_k^{n+1}+u_{k-1}^{n+1}}{h^2}=0,\\ u_k^0=U_0(x_k),\quad k=1,\ 2,\ \ldots,\ K-1,\\ u_0^{n+1}=\alpha^{n+1},\quad u_K^{n+1}=\beta^{n+1}, \end{aligned}\right\}\qquad (189)$$

où α^{n+1}, β^{n+1} sont donnés. Il est évident que le problème (189) donne une approximation du problème différentiel du type (177) sur l'intervalle fini x avec les conditions correspondantes sur ses extrémités.

Pour démontrer la stabilité du problème aux différences (189) il y a lieu de l'écrire sous la forme $lu=f$ et de vérifier que la solution u est du même ordre de grandeur que f, pour f quelconque. Bien que les équations aux différences (189) soient homogènes, il faut ajouter f_k^n dans les deuxièmes membres et montrer que la solution est du même ordre de grandeur que f, α, β, U_0 (la non-homogénéité apparaîtra lors de l'étude de la convergence par suite de l'erreur d'approximation).

Estimons $|u_k^{n+1}|$. Supposons que le maximum de cette grandeur sur la couche corresponde au point k. Alors, si $u_k^{n+1}>0$, on a

$$u_{k+1}^{n+1}-2u_k^{n+1}+u_{k-1}^{n+1}\leqslant 0,$$

les valeurs de gauche et de droite étant majorées par la moyenne. Si $u_k^{n+1} < 0$, la dernière expression $\geqslant 0$. En vertu de (189) et en ajoutant f_k^n, le signe de cette expression se conserve également pour la grandeur

$$\frac{u_k^{n+1} - u_k^n}{\tau} - f_k^n.$$

Par conséquent dans les deux cas

$$|u_k^{n+1}| \leqslant |u_k^n + \tau f_k^n| \leqslant |u_k^n| + \tau |f_k^n|.$$

Par hypothèse, dans le premier membre on a la valeur maximale de $|u_k^{n+1}|$ sur la $(n+1)$-ième couche. En utilisant comme d'habitude la désignation

$$\|u^n\| = \max_k |u_k^n|,$$

on obtient à partir de la dernière inégalité

$$\|u^{n+1}\| \leqslant |u_k^n| + \tau |f_k^n| \leqslant \|u^n\| + \tau \|f^n\|.$$

Si par contre max $|u_k^{n+1}|$ correspond à la limite, il est égal à $|\alpha^{n+1}|$ ou $|\beta^{n+1}|$. Ainsi,

$$\|u^{n+1}\| \leqslant \max(|\alpha^{n+1}|, |\beta^{n+1}|, \|u^n\| + \tau \|f^n\|).$$

En appliquant cette inégalité pour l'estimation ultérieure de u^n à l'aide de u^{n-1}, etc., on obtient

$$\|u^{n+1}\| \leqslant \max(\|\alpha\|, \|\beta\|, \|U_0\| + (n+1)\tau \|f\|) \quad (190)$$

avec les notations suivantes

$$\|f\| = \max_n \|f^n\|, \qquad \|\alpha\| = \max_n |\alpha^n|, \ldots$$

Comme $(n+1)\tau \leqslant T = \text{const}$, l'estimation (190) signifie que le problème aux différences (189) est stable. Remarquons que le résultat obtenu antérieurement à l'aide du critère spectral s'est trouvé confirmé. Le schéma aux diffé-

rences (189) est stable *quelles que soient* les relations entre les pas τ, h. Du point de vue du domaine de dépendance ce résultat n'a rien d'étonnant. Pour le calcul de chacun des u_k^{n+1} il faut résoudre le système d'équations (189) ; ceci étant, chacun des u_k^n sera pris formellement en considération.

Pour d'autres problèmes liés à l'intégration des équations d'évolution et des systèmes de telles équations, on peut également trouver des schémas aux différences implicites absolument stables. La possibilité de choisir les pas τ, h conformément à la précision requise permet souvent de réduire notablement les calculs, même s'il s'agit de l'augmentation de la quantité d'opérations arithmétiques pour chaque point de calcul due à la nécessité de résoudre des systèmes d'équations.

Comme nous allons le voir ci-dessous, les particularités de ces systèmes permettent lorsque ceux-ci sont linéaires d'utiliser des méthodes de résolution efficaces et relativement simples. Donc lors de l'approximation des équations différentielles non linéaires par des schémas aux différences implicites ces derniers doivent être linéaires par rapport aux grandeurs de la $(n + 1)$-ième couche inconnue.

Supposons que le problème initial soit *quasi linéaire*, c'est-à-dire linéaire par rapport aux dérivées d'ordre élevé. Dans ce cas, en utilisant une approximation implicite pour ces dernières, on obtient un problème aux différences dont les conditions de stabilité seront déterminées par l'approximation explicite des termes et des coefficients d'ordre inférieur. Généralement ces conditions ne sont pas difficiles à remplir.

Donc si, dans le problème envisagé, le coefficient de conductibilité thermique a est une fonction de U, $a = a(U)$, en faisant appel à l'approximation (188) avec $a = a(u_k^n)$, on obtient un schéma aux différences implicite, toujours linéaire par rapport aux inconnues u_k^{n+1}. Le problème aux différences (188) devient bien entendu non linéaire, tout comme l'équation différentielle initiale. En appliquant

les méthodes usuelles (§§ 5, 6), on peut faire apparaître que la limitation imposée aux pas du réseau, à savoir la condition (179), dans ce cas également se trouve levée.

Il ne faut pas penser que l'existence de schémas implicites rend unutiles les schémas explicites. La simplicité et la compacité de ces derniers sont souvent très précieuses, surtout dans le cas des problèmes non linéaires compliqués.

Problèmes

Elaborer et étudier différents schémas aux différences implicites permettant d'étudier dans l'intervalle $0 \leqslant x \leqslant 1$ les problèmes suivants:

1. $$\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} = 0, \quad a > 0,$$
$$U(0, x) = U_0(x), \quad U(t, 0) = \alpha(t).$$

2. $$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} = \frac{\partial}{\partial x} U^2 \frac{\partial U}{\partial x},$$
$$U(0, x) = U_0(x), \quad U(t, 0) = \alpha(t), \quad U(t, 1) = \beta(t).$$

3. $$\frac{\partial U}{\partial t} + c^2\frac{\partial V}{\partial x} = 0,$$
$$\frac{\partial V}{\partial t} + \frac{\partial U}{\partial x} = 0,$$
$$U(0, x) = U_0(x), \quad V(0, x) = V_0(x),$$
$$U(t, 0) = \alpha(t), \quad V(t, 1) = \beta(t).$$

§ 9. RÉSOLUTION DES ÉQUATIONS AUX DIFFÉRENCES

Lorsque l'on utilise des schémas aux différences implicites il faut sur chaque couche résoudre un système d'équations différentielles. Ce n'est qu'après avoir donné la méthode

7—01279

de leur résolution que l'on peut considérer la description de l'algorithme de calcul terminée.

Le nombre d'équations et d'inconnues est très grand, de l'ordre de $1/h$. Si l'on ne tient pas compte du caractère spécifique du système, en le résolvant comme un système du type général, un grand nombre d'opérations arithmétiques est nécessaire, ce nombre est bien plus grand que dans le cas des schémas explicites. On peut montrer que pour les systèmes linéaires le nombre d'opérations nécessaires est de l'ordre de $1/h^3$.

De plus, il ne faut pas oublier qu'un algorithme ne peut être réalisé avec une précision absolue, les calculs étant toujours menés avec un nombre limité de décimales. Lorsque l'ordre du système est élevé, l'accumulation de l'erreur d'arrondi peut provoquer une catastrophe.

Revenons au système linéaire (189) que nous écrivons comme suit

$$\left.\begin{array}{r} u_0 = \alpha, \\ -ru_{k-1} + (1+2r)\,u_k - ru_{k+1} = u_k^n, \\ k = 1, 2, \ldots, K-1, \\ u_K = \beta. \end{array}\right\} \qquad (191)$$

Nous avons omis l'indice $n+1$ auprès des inconnues et introduit la désignation suivante

$$r = a\frac{\tau}{h^2}.$$

Comme nous l'avons vu lors de la démonstration de la stabilité du problème (189), la solution du système (191) doit satisfaire à l'inégalité

$$|u_k| \leqslant \max\left(|\alpha|,\ |\beta|,\ \max_k |u_k^n|\right).$$

Il en découle que le système homogène correspondant, obtenu pour $\alpha = \beta = u_k^n = 0$, n'a qu'une solution triviale $u_k = 0$.

Par conséquent, la solution du système d'équations (191) existe et elle est unique.

Le caractère spécifique du système (191) consiste en ce que chaque k-ième équation ne contient que trois inconnues u_{k-1}, u_k, u_{k+1}. Ceci permet d'exclure successivement u_0, u_1, u_2, . . . par la méthode simple suivante. La valeur u_0 est donnée, donc l'équation correspondant à $k = 1$ ne contient en fait que deux inconnues u_1 et u_2, c'est-à-dire qu'elle donne la relation existant entre elles. A l'aide de cette relation, en utilisant l'équation suivante, on exclut u_1 et on obtient une relation entre u_2 et u_3, etc. Considérons la relation existant entre u_{k-1} et u_k:

$$u_{k-1} = L_k u_k + M_k. \tag{192}$$

Substituons cette expression dans la k-ième équation et résolvons celle-ci par rapport à u_k. On obtient

$$u_k = \frac{ru_{k+1} + u_k^n + rM_k}{1 + 2r - rL_k},$$

c'est-à-dire la relation existant entre la paire suivante d'inconnues u_k et u_{k+1}. Ecrivons-la sous la forme (192) posant à cet effet

$$L_{k+1} = \frac{r}{1 + 2r - rL_k}, \tag{193}$$

$$M_{k+1} = \frac{u_k^n + rM_k}{1 + 2r - rL_k}. \tag{194}$$

Ces deux formules permettent de passer de L_k, M_k à L_{k+1}, M_{k+1} pour k quelconque. Comme $u_0 = \alpha$, en vertu de (192) il y a lieu de poser $L_1 = 0$, $M_1 = \alpha$ et à l'aide des formules de récurrence (193), (194) calculer successivement tous les L_k, M_k jusqu'à L_K, M_K inclus. Puis, comme $u_K = \beta$, en utilisant la formule (192), on trouve successivement tous les u_k.

Ainsi, le processus de résolution du système d'équations linéaires (191) se réduit au calcul successif, à l'aide des

formules (193), (194), des coefficients L_k, M_k et suivi du calcul de u_k à l'aide de la formule (192). C'est pourquoi on appelle ce processus *méthode du balayage*. L'avantage essentiel de cette méthode est d'être très économique. Il est facile de voir que le nombre d'opérations arithmétiques exigé par cette méthode est du même ordre de grandeur que le nombre d'inconnues, soit $\sim 1/h$, c'est-à-dire est minimal.

Nous allons maintenant vérifier la sensibilité de la méthode du balayage aux erreurs d'arrondi, c'est-à-dire trouver la relation existant entre la précision des calculs et la précision de la solution obtenue. Pour estimer le développement et l'accumulation de ces erreurs nous allons supposer que le calcul approché (avec arrondissement) à l'aide des formules (192), (193), (194) peut être interprété comme un calcul exact suivant d'autres formules voisines de celles-ci.

Nous allons commencer par la formule (193) donnant la transition de L_k à L_{k+1}. L'erreur de la valeur effectivement calculée L_{k+1} par rapport à sa valeur exacte apparaît pour deux raisons. Primo, la valeur utilisée L_k contient l'erreur δL_k provenant des calculs antérieurs. Pour trouver son influence sur l'erreur δL_{k+1}, on prend la dérivée de l'expression (193) par rapport à L_k. On obtient :

$$\delta L_{k+1} = (L_{k+1})^2 \delta L_k.$$

Secundo, comme le résultat de chaque opération arithmétique est arrondi, même si la valeur utilisée L_k est exacte, L_{k+1} sera entaché de l'erreur apparaissant dans le cycle donné de calcul. Désignons cette erreur par δ_k.

Ainsi, en tous cas pour $\delta L \ll L$, on a

$$\delta L_{k+1} = L_{k+1}^2 \delta L_k + \delta_k. \tag{195}$$

La grandeur δ_k caractérise la précision des calculs.

On voit donc que l'évolution de l'erreur d'arrondi se trouve entièrement déterminée par les coefficients L_k. En particulier, si $| L_k | > 1$, l'augmentation de δL_k est expo-

nentielle, la précision décroissant rapidement. Nous allons donner une estimation de la grandeur L_k.

Comme $L_1 = 0$, en vertu de (193), on a

$$L_2 = \frac{r}{1+2r} < 1.$$

Supposons que $0 < L_k < 1$. Il est facile de voir, à partir de (193), qu'alors on a l'estimation suivante

$$0 < L_{k+1} < \frac{r}{1+r} < 1, \tag{196}$$

c'est-à-dire que tous les L_k ne sont pas supérieurs à $r/(1+r)$.

En utilisant la formule de récurrence (195), on obtient

$$|\delta L_N| \leqslant |\delta_{N-1}| + \left(\frac{r}{1+r}\right)^2 |\delta L_{N-1}| \leqslant$$
$$\leqslant |\delta_{N-1}| + \left(\frac{r}{1+r}\right)^2 |\delta_{N-2}| + \left(\frac{r}{1+r}\right)^4 |\delta L_{N-2}| \leqslant$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\leqslant \max_k |\delta_k| \frac{1}{1-\left(\frac{r}{1+r}\right)^2} + \left(\frac{r}{1+r}\right)^{2N} |\delta L_0|,$$

ce qui montre immédiatement que l'erreur δL est de l'ordre de δ.

Les formules (194) et (192) peuvent être étudiées d'une manière analogue. En dérivant, on obtient

$$\delta M_{k+1} = L_{k+1}\delta M_k + \delta'_k,$$
$$\delta u_{k-1} = L_k \delta u_k + \delta''_k,$$

où δ', δ'' contiennent les erreurs δu^n, $\delta\alpha$, $\delta\beta$. En utilisant l'inégalité (196), on voit que lors du calcul de M et u l'accumulation de l'erreur d'arrondi n'a pas lieu non plus.

Nous avons montré que lorsqu'on résout le système d'équations linéaires (191) par la méthode du balayage, la précision du résultat *coïncide* avec celle des calculs et des données initiales.

Cette affirmation n'est pas vraie pour toutes les méthodes. Nous allons voir un exemple. En écrivant chacune des équations du système (191) sous la forme suivante

$$u_{k+1} = -u_{k-1} + \frac{1+2r}{r} u_k - \frac{1}{r} u_k^n, \tag{197}$$

on peut, à partir de la paire de valeurs u_{k-1}, u_k, calculer u_{k+1}. Pour commencer le processus, on a besoin des valeurs u_0 et u_1. Pour la première d'entre elles on a $u_0 = \alpha$. La seconde peut prendre une valeur arbitraire, $u_1 = 0$ par exemple. Les autres u_k peuvent être calculées à partir de (197). Désignons par $u_k^{(1)}$ la solution obtenue. Il est évident que dans la pratique elle ne satisfera pas à la condition limite $u_K = \beta$. Trouvons la seconde solution $u_k^{(2)}$ de la même manière en posant $u_1^{(2)} = 1$. Considérons une combinaison linéaire de ces solutions

$$u_k = cu_k^{(1)} + (1-c)\, u_k^{(2)} \tag{198}$$

qui de toute évidence satisfait à la condition limite à l'extrémité gauche et aux équations. Nous allons trouver c de façon à satisfaire à la condition limite $u_K = \beta$. On a

$$c = \frac{\beta - u_K^{(2)}}{u_K^{(1)} - u_K^{(2)}}\,. \tag{199}$$

Alors la formule (198) donne la solution du problème.

A première vue, la méthode décrite est même plus commode que la méthode du balayage. Cependant elle ne peut être utilisée ce que nous allons montrer sur un cas particulier où l'on peut écrire la solution sous forme explicite. Posons $u_k^n = 0$ et cherchons la solution sous la forme $u_k = \text{const}\cdot q^k$. En substituant cette expression dans (197), on obtient une équation quadratique en q, soit:

$$q^2 - \left(2 + \frac{1}{r}\right) q + 1 = 0.$$

Pour $r > 0$ quelconque ses racines sont réelles, et ce qui est important, l'une d'elles est supérieure à l'unité, $q_1 < 1$,

$q_2 > 1$. Toute solution du problème est une combinaison linéaire de q_1^k et q_2^k, il est alors facile de montrer que dans le cas présent $u_k^{(1)}$ et $u_k^{(2)}$ mentionnés ci-dessus s'écrivent

$$u_k^{(1)} = \frac{\alpha q_2 q_1^k - \alpha q_1 q_2^k}{q_2 - q_1}, \quad u_k^{(2)} = \frac{(\alpha q_2 - 1) q_1^k + (1 - \alpha q_1) q_2^k}{q_2 - q_1},$$

par conséquent, la solution de (198) est

$$u_k = u_k^{(2)} - c\,\frac{q_2^k - q_1^k}{q_2 - q_1}. \tag{200}$$

Nous allons voir les conséquences des erreurs d'arrondi lorsqu'on réalise le processus (197)-(199). Même en idéalisant et en supposant que l'on arrondisse seulement la grandeur c de (199), tous les autres calculs étant faits avec une précision absolue, la solution sera quand même entachée d'une erreur égale à

$$\delta u_k \sim \delta c\,\frac{q_2^k - q_1^k}{q_2 - q_1},$$

ceci en vertu de (200). Comme $q_2^k \gg 1$ pour k grands, une petite erreur dans le calcul de c entraîne une grosse erreur dans la solution. La grandeur k est de l'ordre de $1/h$. Ainsi, pour atteindre une certaine précision dans la solution, les calculs doivent être menés avec une précision $q_2^{1/h}$ fois plus grande. Par exemple pour $q_2 = 2$ et $k \sim 100$ on obtient $q_2^k \sim 2^{100} \sim 10^{30}$. Il est évident que l'on ne peut avoir une telle réserve de décimales. C'est pourquoi cette méthode correcte du point de vue théorique doit être rejetée.

Nous avons exposé la méthode du balayage sur l'exemple du système (191) correspondant à l'équation de la conductibilité thermique pour l'intervalle fini $(0, X)$ avec les conditions aux limites $U(t, 0) = \alpha(t)$, $U(t, X) = \beta(t)$. Pour que le lecteur puisse se faire une idée des possibilités de généralisation de la méthode exposée à des problèmes plus compliqués, nous allons nous arrêter sur certaines questions fondamentales.

Toute condition aux limites peut être interprétée comme l'expression effective de l'influence des processus ayant lieu dans la partie extérieure, rejetée de l'espace. De ce point de vue les équations différentielles décrivent elles-mêmes la diffusion de cette influence à l'intérieur du domaine. Le problème aux différences est une approximation du problème différentiel. En particulier on peut lui appliquer également les considérations sur le rôle des conditions aux limites et des équations. La méthode du balayage exprime d'une manière explicite ce processus de diffusion de l'influence des conditions aux limites à l'intérieur du domaine dans le cas où les équations différentielles sont linéaires.

Le schéma général de la méthode du balayage pour un problème linéaire quelconque est le même que celui qui a été envisagé pour les cas de la conductibilité thermique. La réalisation aux différences des conditions aux limites sur l'une des extrémités de l'intervalle de calcul (par exemple sur celle de gauche) donne les relations entre les valeurs de la fonction discrète (ou du vecteur fonction) aux points limite ou voisin de la limite. En utilisant les équations aux différences on exclut successivement les inconnues, on fait « parcourir » la relation entre les valeurs voisines de la fonction discrète à travers tout le domaine de calcul jusqu'à l'extrémité droite. Les relations entre les conditions aux limites que l'on a ici permettent de trouver la valeur de la fonction au dernier point. En se déplaçant maintenant dans le sens inverse, à l'aide des relations obtenues antérieurement, on trouve successivement toutes les valeurs de la fonction discrète.

Evidemment ce n'est là que le schéma de la méthode, sa forme concrète pouvant changer notablement d'un problème à l'autre, suivant les particularités des conditions aux limites et des équations.

Pour la solution des systèmes d'équations aux différences on peut également utiliser d'autres méthodes, en particulier *itératives*. Revenons au système d'équations (191) et

récrivons-le sous la forme suivante:

$$\left.\begin{aligned} u_0 &= \alpha, \\ u_k &= \frac{r(u_{k-1}+u_{k+1})+u_k^n}{1+2r}, \quad k=1,\ 2,\ \ldots,\ K-1, \\ u_K &= \beta, \end{aligned}\right\} \tag{201}$$

que nous allons utiliser pour le processus itératif. En substituant dans le second membre la ν-ième approximation de $u_k^{(\nu)}$, on obtient $u_k^{(\nu+1)}$ dans le premier membre. Etudions la convergence des itérations. Posons

$$u_k^{(\nu)} = u_k + \delta_k^{(\nu)},$$

où u_k est la solution exacte du système (201). Il est facile d'obtenir les formules suivantes:

$$\begin{aligned} &\delta_0^{(\nu+1)} = 0, \\ &\delta_k^{(\nu+1)} = \frac{r}{1+2r}\left(\delta_{k-1}^{(\nu)} + \delta_{k+1}^{(\nu)}\right), \quad k=1,\ 2,\ \ldots,\ K-1, \\ &\delta_K^{(\nu+1)} = 0, \end{aligned}$$

déterminant l'erreur en fonction du numéro de l'itération. L'estimation directe de $|\delta_k^{(\nu+1)}|$ donne

$$\max_k |\delta_k^{(\nu+1)}| \leqslant \frac{2r}{1+2r} \max_k |\delta_k^{(\nu)}|.$$

Par conséquent, les itérations convergent, $\delta_k^{(\nu)} \to 0$, la rapidité de convergence étant

$$\Theta = \frac{2r}{1+2r} < 1.$$

Arrêtons-nous sur le nombre d'itérations. Supposons que l'on prolonge le processus jusqu'à la coïncidence de deux itérations successives $u_k^{(\nu+1)} = u_k^{(\nu)}$, c'est-à-dire que l'on obtient la solution du système (201) avec une précision maximale. Cependant une telle précision n'est pas indis-

pensable car le système (201) lui-même n'est qu'une approximation du problème initial différentiel. Lorsque l'on utilise la méthode du balayage on obtient cette précision gratuitement, ici au contraire pour l'obtenir il faut faire appel à des itérations successives, par un travail supplémentaire pouvant par conséquent être superflu.

D'un autre côté, la convergence de la méthode numérique (découlant de l'approximation et de la stabilité) a été démontrée en supposant que le problème aux différences était exact. Il est facile de montrer que si l'on tient compte des erreurs d'arrondi, cela ne change rien (car la stabilité les amortit). L'influence des erreurs apparaissant par suite des itérations insuffisantes doit être étudiée spécialement.

En exagérant, supposons que l'on se borne à une seule itération en prenant pour l'approximation initiale $u_k^{(0)} = u_k^n$. Ceci signifie qu'en fait pour le calcul de u_k^{n+1}, on utilise la formule suivante [voir (201)]:

$$u_k^{n+1} = \frac{r(u_{k-1}^n + u_{k+1}^n) + u_k^n}{1+2r}. \tag{202}$$

En estimant $|u_k^{n+1}|$ on obtient

$$\max_k |u_k^{n+1}| \leqslant \max_k |u_k^n|,$$

c'est-à-dire le schéma aux différences (202) est stable (pour r quelconque, bien que le schéma soit explicite!).

Revenons à l'approximation. En substituant dans (202) les développements suivants

$$U_k^{n+1} = U + U_t\tau + O(\tau^2),$$

$$U_{k\pm 1}^n = U \pm U_x h + U_{xx}\frac{h^2}{2} \pm U_{xxx}\frac{h^3}{6} + O(h^4),$$

on obtient

$$U_t + O(\tau) = \frac{aU_{xx} + O(h^2)}{1+2r}.$$

Par conséquent, la relation aux différences (202) ne donne une approximation de l'équation de la conductibilité thermi-

que que pour $r \to 0$. Si r par contre est fini, en utilisant (202) on aura une autre équation. On peut montrer (voir problème 4) qu'il en sera de même après la ν-ième itération, c'est-à-dire que l'erreur d'approximation est une fonction de r et de ν et ne tend vers zéro que pour $r \to 0$ ou $\nu \to \infty$. Ainsi, un nombre insuffisant d'itérations peut donner lieu à l'absence d'approximation.

Un autre cas est possible. Supposons, par exemple, que pour la solution du problème (191) au lieu de (201) on ait un autre processus itératif du type

$$u_k^{(\nu+1)} = \varphi\left(u_{k-1}^{(\nu)},\ u_k^{(\nu)},\ u_{k+1}^{(\nu)},\ u_k^n\right),$$

pour lequel la condition d'approximation se trouve vérifiée. Il est évident que dans le cas présent on risque de ne pas vérifier la condition de stabilité. Pour estimer le nombre nécessaire d'itérations nous allons raisonner de la manière suivante. Lors de la première itération, pour trouver $u_k^{(1)}$ on tient compte en plus de u_k^n seulement de $u_{k-1}^n = u_{k-1}^{(0)}$ et de $u_{k+1}^n = u_{k+1}^{(0)}$. Lors de la seconde, en utilisant $u_{k-1}^{(1)}$ et $u_{k+1}^{(1)}$ on tient compte par là même de $u_{k-2}^n = u_{k-2}^{(0)}$ et $u_{k+2}^n = u_{k+2}^{(0)}$, etc. Chaque itération étend le domaine de dépendance d'un point de calcul, ceci d'un côté et de l'autre. C'est pourquoi ν itérations sont équivalentes à un certain schéma explicite utilisant $2\nu + 1$ valeurs de la n-ième couche. Pour la stabilité de ce dernier, il faut vérifier une certaine inégalité du type

$$\frac{\tau}{(\nu h)^2} \leqslant \text{const},$$

c'est-à-dire $r \leqslant \text{const} \cdot \nu^2$. Ainsi le nombre nécessaire d'itérations est égal à

$$\nu \sim \sqrt{r}.$$

On voit que lorsque r n'est pas trop grand, les méthodes itératives peuvent donner des résultats satisfaisants et même faire concurrence à la méthode du balayage car elles per-

mettent de diviser le processus de calcul en plusieurs branches parallèles et n'ont pas besoin d'une grande mémoire fixant des massifs importants de coefficients L, M indispensables pour la méthode du balayage. Il est parfois commode de combiner les deux méthodes en utilisant le balayage sur les domaines isolés et de les « ajuster » par itérations.

Nous allons faire une dernière remarque. Nous n'avons étudié que les équations aux différences linéaires. Les méthodes itératives ont un champ d'application plus large, par contre la méthode du balayage utilise en fait la linéarité des équations. Pour la résolution des équations non linéaires seules les méthodes itératives conviennent. Cependant les itérations peuvent être réalisées différemment. Ci-dessus nous avons envisagé un algorithme explicite du type $u^{(\nu+1)} = \varphi(u^{(\nu)})$. Pour r grand il converge lentement car il faut tenir compte du domaine de dépendance de la solution. Ce dernier est donné par les propriétés du problème, celles-ci apparaissant déjà dans son modèle linéaire. Ayant à notre disposition un instrument aussi efficace que la méthode du balayage, on peut lors de l'élaboration du processus itératif utiliser des algorithmes implicites linéaires et ramener ainsi le problème à la résolution d'un système linéaire sur chaque itération. Ceci permet d'obtenir un processus à convergence rapide.

Supposons, par exemple, que dans le problème envisagé ci-dessus le coefficient a dépende de la fonction cherchée $a = a(U)$. Il est évident que lors de l'approximation par un problème aux différences on peut prendre $a(u^n)$, mais supposons que pour des raisons quelconques ceci ne nous convienne pas et que l'on utilise $a(u^{n+1})$. Le système d'équations aux différences (191) se trouve alors être non linéaire $r = r(u_k)$. Il est tout naturel de poser $r = r(u_k^{(\nu)})$ et en résolvant le système (191) par la méthode du balayage d'obtenir $u_k^{(\nu+1)}$. Si pour l'approximation initiale on prend $u_k^{(0)} = u_k^n$, la rapidité de convergence du processus itératif

$u_k^{(\nu)} \to u_k^{n+1}$ sera déterminée par la grandeur de la variation effective de la solution d'une couche à l'autre.

Problèmes

1. Estimer l'ordre de grandeur du nombre d'opérations arithmétiques nécessaires à la résolution du système de N équations linéaires du type général par la méthode d'exclusion.

2. Si la méthode de résolution d'un système d'équations obtenu à partir d'un schéma implicite est donnée, l'opérateur de transition d'une couche à l'autre se trouve ainsi déterminé, on peut alors étudier le schéma en tant qu'explicite et l'écrire comme suit

$$u^{n+1} = R(u^n + \tau f^n).$$

En utilisant la stabilité du problème, c'est-à-dire la condition limitant les puissances de l'opérateur R, $\| R^n \| \leqslant \text{const}$, étudier le comportement des erreurs apparaissant dans u^{n+1} par suite de l'arrondissement du second membre lors de l'augmentation de n, $n\tau < t$.

3. Etudier l'équation aux différences

$$u_0^{n+1} = \alpha^{n+1},$$

$$\left.\begin{aligned} \frac{u_k^{n+1} - u_k^n}{\tau} + \frac{v_{k+1/2}^{n+1} - v_{k-1/2}^{n+1}}{h} &= 0, \\ \frac{v_{k-1/2}^{n+1} - v_{k-1/2}^{n}}{\tau} + \frac{u_k^{n+1} - u_{k-1}^{n+1}}{h} &= 0, \end{aligned}\right\} k = 1, 2, \ldots, K,$$

$$v_{K+1/2}^{n+1} = \beta^{n+1}$$

où $v_{k+1/2}$ signifie que cette valeur correspond au point $x_{k+1/2} = (k + 1/2)\, h$. Appliquer la méthode du balayage en utilisant une relation du type

$$u_{k-1} = L_k v_{k-1/2} + M_k, \quad \text{ou} \quad u_{k-1} = L_k u_k + M_k.$$

Vérifier si l'algorithme de calcul est correct par rapport aux erreurs d'arrondi.

Trouver le schéma analogue pour la résolution du problème non linéaire suivant

$$\frac{\partial U}{\partial t}+\frac{\partial P(V)}{\partial x}=0, \qquad U(0, x)=U_0(x),$$

$$\frac{\partial V}{\partial t}+\frac{\partial U}{\partial x}=0, \qquad V(0, x)=V_0(x)$$

sur l'intervalle $0 \leqslant x \leqslant X$ sur les bords duquel on a les conditions aux limites suivantes

$$U(t, 0) = \alpha(t), \quad V(t, X) = \beta(t).$$

4. Supposons que, lors de la réalisation du processus étudié ci-dessus (201)

$$u_k^{(\nu+1)}=\frac{r\left(u_{k-1}^{(\nu)}+u_{k+1}^{(\nu)}+u_k^n\right)}{1+2r}, \quad k=0, \pm 1, \ldots,$$

on se borne à N itérations, c'est-à-dire $u^{n+1} = u_k^{(N)}$, $u_k^{(0)} = u_k^n$ (pour plus de simplicité omettons les conditions aux limites).

Vérifier que le schéma aux différences obtenu est stable pour r et N quelconques.

Montrer que l'erreur d'approximation est proportionnelle à

$$\left(\frac{2r}{1-2r}\right)^N.$$

A cet effet poser

$$u_k^{(\nu)}=a^{(\nu)}+hb^{(\nu)}k+\frac{h^2}{2}c^{(\nu)}k^2+\ldots,$$

où de toute évidence

$$a^{(0)}=u_0^n, \ b^{(0)}=(u_x)_0^n, \ c^{(0)}=(u_{xx})_0^n,$$

et trouver $a^{(N)}$ à l'aide de la formule du processus itératif. Comme

$$u_0^{n+1}=a^{(N)}=u_0^n+\tau(u_t)_0^n+\ldots,$$

connaissant $a^{(N)}$ il est facile de trouver l'erreur d'approximation.

Chapitre III

§ 10. CALCUL DES SOLUTIONS DISCONTINUES

Dans notre étude antérieure nous avons supposé que la solution exacte du problème initial est une fonction régulière. Lors de l'étude de l'approximation et de l'élaboration des modèles linéaires nous avons souvent utilisé cette hypothèse. Pour les problèmes différentiels c'est tout naturel, la solution devant être telle qu'au moins les dérivées figurant dans l'équation existent. Si la solution contient des singularités la rendant non régulière, il y a lieu de procéder à une étude spéciale et en cas de nécessité de modifier la méthode. On ne peut rien dire de plus précis, chaque singularité ayant ses particularités.

Ici nous nous bornerons à l'étude de cette question dans le cas important pour la pratique où la solution est une fonction lisse par morceaux avec discontinuités. Considérons l'exemple

$$\frac{\partial U}{\partial t}+U\frac{\partial U}{\partial x}=0, \qquad U(0,\ x)=U_0(x). \tag{203}$$

Dans le cas où la solution $U\ (t,\ x)$ est une fonction régulière, on est ramené au problème que nous avons déjà étudié en détail. Cependant dans certains cas les conditions de régularité et d'existence de la solution peuvent s'exclure mutuellement. Ainsi la solution du problème (203) satisfait évidemment au système

$$\frac{dx}{dt}=U, \qquad \frac{dU}{dt}=0,$$

c'est-à-dire qu'elle est constante le long des droites

$$x=x_0+U_0\,(x_0)\,t.$$

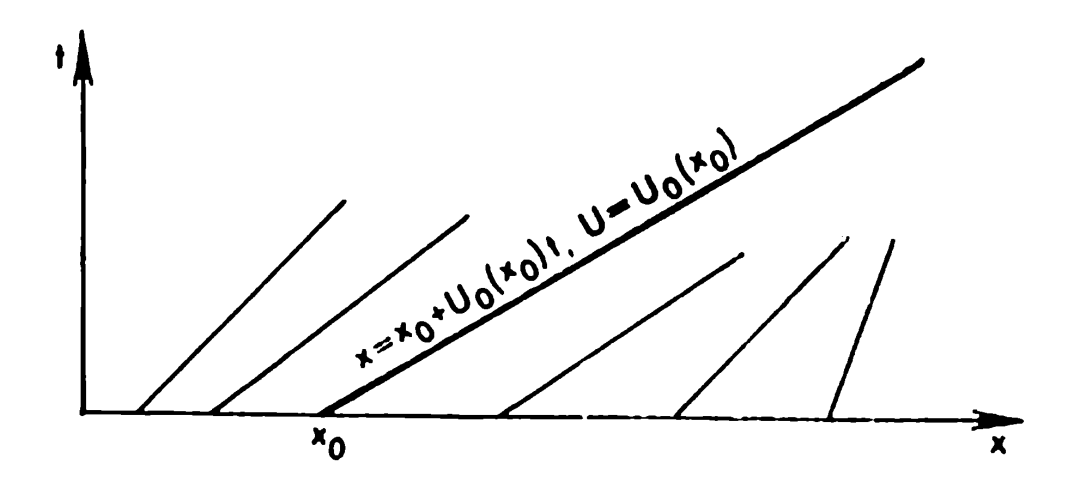


Fig. 12

Ces lignes sont appelées *caractéristiques*. Le long de ces caractéristiques les équations du problème dégénèrent en relations entre les différentielles de la fonction (ou des fonctions). Dans notre problème la pente de chacune des caractéristiques est donnée par la valeur de la fonction $U_0(x)$ au point x_0, point d'intersection d'une caractéristique considérée avec la ligne des données initiales $t = 0$ (fig. 12). Il est facile de voir que dans le cas où $U_0(x)$ est, ne serait-ce que sur un petit segment de l'axe x, une fonction décroissante, les caractéristiques issues des points de ce segment se rencontrent. Chacune d'elles donnant indépendamment des autres une valeur de la fonction, la solution au point d'intersection ne sera pas univoque. Pour parer à ces difficultés, pour les problèmes décrivant des processus physiques réels (hydrodynamique), on admet l'existence de solutions discontinues.

En particulier la non-unicité mentionnée indiquera l'apparition d'une discontinuité. Mais si l'on suppose que la solution contient une discontinuité, au point de discontinuité les dérivées ne sont pas déterminées et par conséquent les équations différentielles perdent leur sens. Nous devons alors remplacer ces dernières par les relations finies reliant les valeurs des fonctions de part et d'autre de la discontinuité. En prolongeant l'analogie existant avec les problèmes hydrodynamiques, il y a lieu de supposer que ces relations

doivent exprimer, pour les solutions discontinues, les mêmes relations physiques que des équations différentielles pour les solutions lisses. Du point de vue mathématique formel ces solutions s'obtiennent de la manière suivante.

Soit $U(x, t)$ une fonction lisse satisfaisant dans un certain domaine du plan x, t à l'équation (203). Intégrons (203) sur une partie quelconque S de ce domaine. Il est facile d'obtenir

$$\iint_S \left(\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x}\right) dx\,dt = \iint_S \left(\frac{\partial U}{\partial t} + \frac{\partial}{\partial x}\left(\frac{U^2}{2}\right)\right) dx\,dt =$$

$$= \iint_S \frac{\partial U}{\partial t}\,dt\,dx + \iint_S \frac{\partial}{\partial x}\left(\frac{U^2}{2}\right) dx\,dt = \oint_\Gamma U\,dx - \frac{U^2}{2}\,dt,$$

où la dernière intégrale curviligne est prise sur le contour Γ, frontière de S. Par conséquent si $U(x, t)$ est la solution de (203), on a

$$\oint_\Gamma U\,dx - \frac{U^2}{2}\,dt = 0 \tag{204}$$

pour un contour quelconque Γ. Lorsque l'on utilise (204) au lieu de (203), la fonction $U(x, t)$ n'a pas besoin d'être ni lisse, ni continue. C'est pourquoi il y a lieu de l'utiliser pour obtenir des relations sur la discontinuité.

Nous allons raisonner comme suit. La discontinuité dans la solution une fois apparue se déplace ensuite décrivant dans le plan x, t une certaine courbe $x = X(t)$. Considérons un petit élément de cette courbe et construisons sur cet élément comme diagonale le contour rectangulaire Γ (fig. 13). Comme ce rectangle est petit, sur chacune de ces deux moitiés on peut supposer que la fonction U est constante et égale à U^- à gauche et U^+ à droite. Calculons l'intégrale (204) pour ce contour, on obtient

$$U^+\Delta x - \left(\frac{U^2}{2}\right)^+ \Delta t - U^-\Delta x + \left(\frac{U^2}{2}\right)^- \Delta t = 0,$$

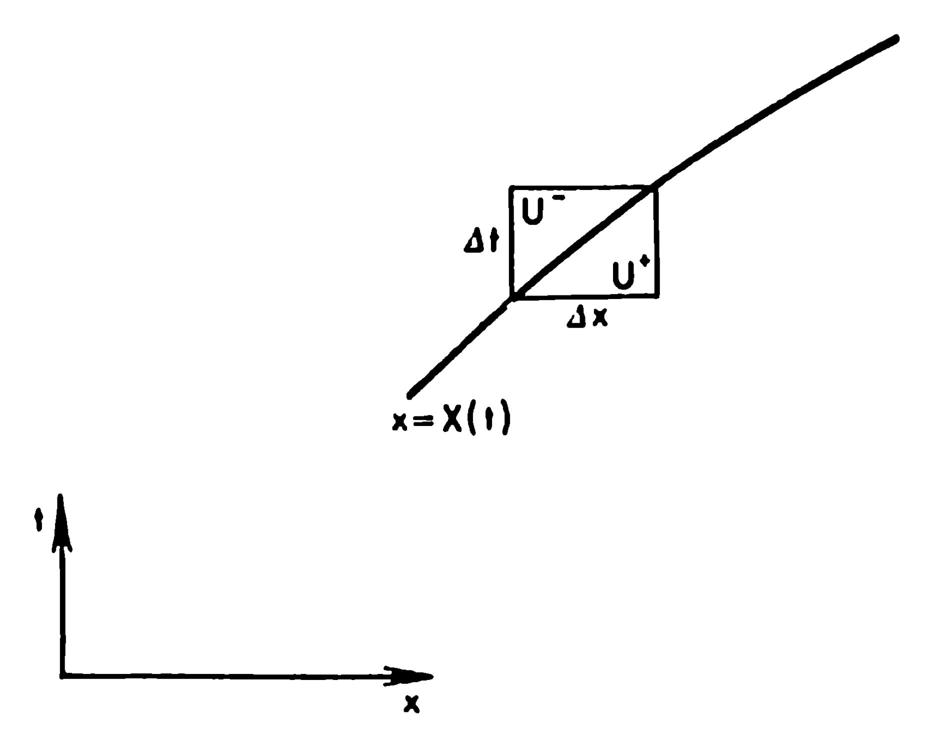


Fig. 13

où Δx, Δt sont les côtés du rectangle. Lorsque l'on contracte le rectangle en un point, le rapport $\Delta x/\Delta t$ tend vers $X'(t)$, à la limite la dernière égalité donne

$$(U^+ - U^-)\,X' = \left(\frac{U^2}{2}\right)^+ - \left(\frac{U^2}{2}\right)^- . \qquad (205)$$

Comme nous l'avons vu ci-dessus, l'intersection des caractéristiques et l'apparition de la discontinuité n'ont lieu que si $U^- > U^+$. On peut montrer que (pour notre problème) seules des discontinuités de ce type peuvent exister. En simplifiant (205) par $U^+ - U^-$ et en y ajoutant l'inégalité mentionnée, on obtient

$$X' = \frac{U^- + U^+}{2}, \qquad U^- > U^+. \qquad (206)$$

Cette relation entre U^-, U^+ et X' remplace sur la ligne de discontinuité $x = X\,(t)$ l'équation différentielle (203).

Généralisons le problème (203) et supposons qu'il y ait des discontinuités sur certaines lignes $x = X(t)$ satisfaisant aux relations (206).

Nous allons passer à l'élaboration de la méthode numérique de résolution des problèmes (203), (206). Le plus simple est d'utiliser les méthodes étudiées ci-dessus en les modifiant seulement au voisinage immédiat de la ligne de discontinuité. Ce n'est pas difficile, tout au moins pour notre problème simple, cependant un certain nombre d'inconvénients apparaissent. Les formules spéciales pour le calcul des grandeurs sur la ligne de discontinuité et dans les points voisins doivent tenir compte de toutes les dispositions possibles de cette ligne par rapport aux points du réseau de base. De plus il faut envisager l'apparition éventuelle d'une discontinuité et par conséquent vérifier la solution obtenue d'une façon convenable (dans tous les points de calcul). Il est évident que cela augmente notablement le volume de l'algorithme de calcul qui perd alors sa simplicité et sa compacité, vu le nombre restreint de points de calcul. C'est pourquoi souvent on préfère une autre méthode pour obtenir les formules de calcul que nous allons décrire ci-dessous.

Pour le problème aux différences la notion de discontinuité de la solution du point de vue formel n'a pas de sens, la fonction discrète étant déterminée sur un ensemble discret de points, le réseau de calcul. D'un autre côté, comme nous l'avons déjà vu, la condition sur la discontinuité (206) découle de l'équation différentielle (203) et par conséquent y est « contenue ». La théorie des équations différentielles nous apprend que l'on peut obtenir la solution discontinue comme la limite de la solution lisse de l'équation perturbée lorsque le paramètre perturbateur tend vers zéro. Nous pouvons utiliser ce fait car en utilisant telle ou telle méthode numérique on remplace toujours le problème initial par un autre problème discontinu perturbé. Nous allons réaliser l'approximation en deux étapes. Tout d'abord en introduisant

une perturbation, on remplace le problème initial par un problème intermédiaire, puis on passe au problème aux différences.

Au lieu de (203) considérons l'équation

$$\frac{\partial U}{\partial t}+U\frac{\partial U}{\partial x}+\varepsilon^2\frac{\partial}{\partial x}\left(\frac{\partial U}{\partial x}\right)^2=0, \qquad (207)$$

où ε est un paramètre petit. Il est évident que si l'on se borne à des fonctions lisses, alors, ε étant petit, les solutions des problèmes (207) et (203) pour des conditions initiales identiques ou voisines seront également voisines. Pour se faire une idée de la différence apparaissant dans le cas des solutions discontinues (pour (203)), considérons le cas particulier suivant.

Supposons que la solution du problème (203), (206) soit une fonction en escalier

$$U=\begin{cases} U^- & \text{pour} \quad x-\omega t<0, \\ U^+ & \text{pour} \quad x-\omega t>0, \end{cases} \qquad (208)$$

où

$$\omega=\frac{U^-+U^+}{2}, \quad U^->U^+, \qquad (209)$$

et U^-, U^+ sont des constantes. Il est évident que la fonction (208) satisfait à (203), (206).

Cherchons la solution correspondante de l'équation (207) sous la forme

$$U_\varepsilon(x, t) = f(x-\omega t). \qquad (210)$$

Il est naturel d'admettre que

$$f(x)\to U^\pm \quad \text{pour} \quad x\to\pm\infty, \qquad (211)$$

car loin de la discontinuité les solutions U, U_ε sont des fonctions lisses et par conséquent voisines. En substituant (210) dans (207), on obtient une équation différentielle

ordinaire pour f

$$-\omega f' + ff' + \varepsilon^2 (f'^2)' = 0. \qquad (212)$$

Il est facile de voir que

$$f = \omega + \text{const} \cdot \sin \frac{x - \omega t}{\varepsilon \sqrt{2}}$$

est la solution de (212). Comme

$$f = \text{const}$$

satisfait également à l'équation (212), la solution qui nous intéresse est de la forme (fig. 14) :

$$U_\varepsilon = \begin{cases} U^- & \text{pour } \dfrac{x - \omega t}{\varepsilon \sqrt{2}} < -\dfrac{\pi}{2}, \\ \dfrac{U^+ + U^-}{2} + \dfrac{U^+ - U^-}{2} \sin \dfrac{x - \omega t}{\varepsilon \sqrt{2}} & \text{pour } \left| \dfrac{x - \omega t}{\varepsilon \sqrt{2}} \right| < \dfrac{\pi}{2}, \\ U^+ & \text{pour } \dfrac{x - \omega t}{\varepsilon \sqrt{2}} > \dfrac{\pi}{2}. \end{cases}$$

Cette solution est une fonction lisse, au lieu de la discontinuité U^-, U^+ elle a une zone de transition continue de U^- à U^+ dont la largeur est $\varepsilon\pi\sqrt{2}$. Si ε est suffisamment petit, ce domaine est étroit et U_ε est voisin de la solution discontinue (208) du problème initial (203), (206).

Ceci permet de remplacer les problèmes (203), (206) par le problème (207). Il est très important que la solution de ce dernier est une fonction lisse. C'est pourquoi lors de l'élaboration de la méthode numérique on peut utiliser toutes les méthodes étudiées ci-dessus permettant de trouver et d'étudier les équations aux différences. L'introduction d'un terme perturbateur (appelé *viscosité artificielle*) complique les calculs mais en revanche donne la possibilité d'effectuer le calcul à l'aide de formules bien connues sans faire appel à l'étude des particularités et aux épreuves sur l'apparition des discontinuités.

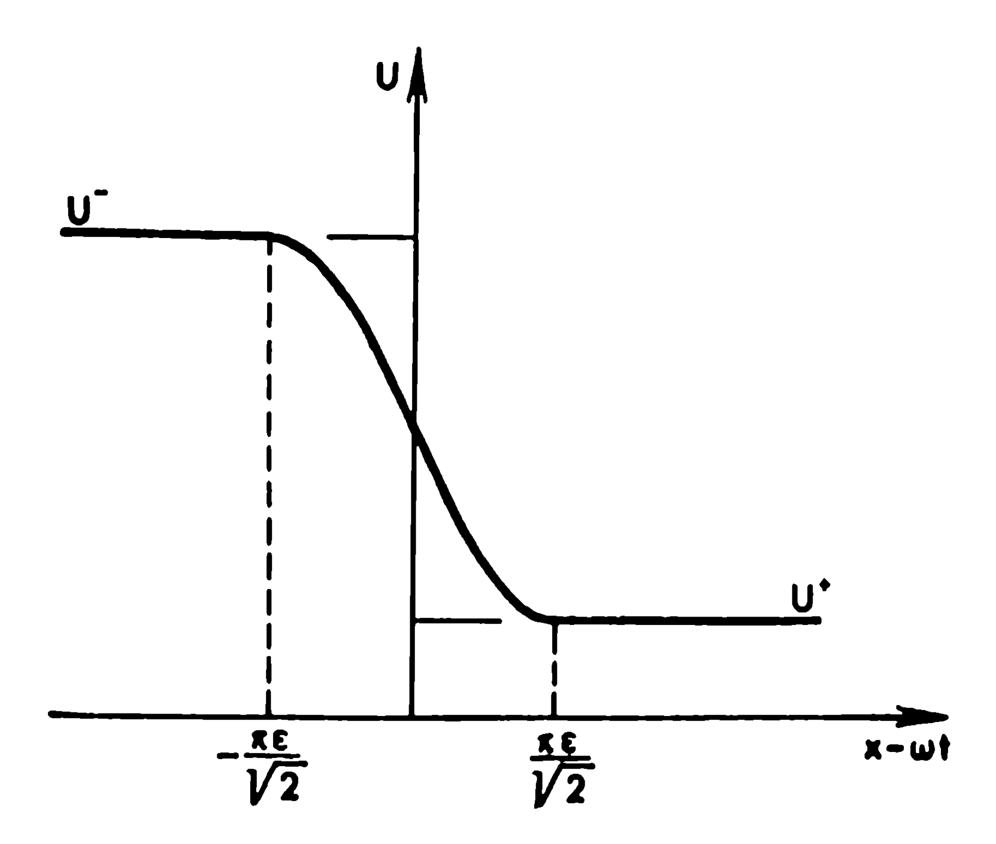


Fig. 14

La méthode envisagée faisant ressortir la viscosité artificielle (207) (qui n'est pas la seule possible) est commode parce que la largeur de la zone « floue » de la discontinuité est de l'ordre de ε et ne dépend pas de la grandeur de la discontinuité $U^- - U^+$. Les variations de la fonction discrète sur des intervalles supérieurs à $\sim\varepsilon$ ont alors un sens réel pour une solution quelconque. C'est donc sur ces considérations que doit être basé le choix de la grandeur ε.

Pour ce qui est du choix de h (et τ pour des schémas implicites) remarquons que l'introduction de la viscosité ne donne l'effet désiré que dans le cas où la zone de discontinuité contient au moins plusieurs points de calcul. Dans le cas contraire il n'y a pas lieu d'attendre une approximation convenable, l'équation aux différences pouvant réagir non correctement à la viscosité.

Ainsi on a une certaine relation entre les paramètres ε, τ, h, c'est-à-dire $\varepsilon \sim \varepsilon(\tau, h)$. D'un autre côté tout problème

aux différences a une certaine viscosité (dite *d'approximation*) car le schéma aux différences est équivalent aux équations initiales majorées de l'erreur d'approximation, par exemple $O(\tau, h)$. Cette dernière est en réalité une certaine expression différentielle à coefficients petits, c'est-à-dire elle peut être interprétée comme une viscosité artificielle. C'est pourquoi, si l'on veut, la méthode exposée ci-dessus d'élaboration du problème aux différences se ramène à la méthode générale avec une erreur d'approximation d'un type spécial.

Nous avons étudié un seul exemple (203) particulièrement simple, néanmoins nous avons exposé tout ce qui est important sur les méthodes de solution de ce type de problèmes lorsqu'il y a des discontinuités, à une exception près.

Bien que ça puisse paraître étrange, ce sont les équations linéaires et en général les problèmes où la discontinuité s'étend le long des caractéristiques (c'est-à-dire la ligne de discontinuité $x = X(t)$ est une caractéristique) qui font exception. Dans ce cas les discontinuités ne peuvent être étalées d'une manière stable à l'aide d'une viscosité artificielle. C'est pourquoi en cas de nécessité, pour le calcul correct d'une telle discontinuité il y a lieu d'utiliser des formules spéciales.

Pour le calcul d'une singularité quelconque deux possibilités s'offrent. Ou bien on donne une description détaillée de la discontinuité ou bien on l'étale. Dans ce paragraphe on a étudié un exemple du second type.

Problèmes

1. Pour les problèmes (203), (206) trouver les formules aux différences pour le calcul des grandeurs sur la discontinuité et aux points voisins. Une maille de calcul typique est donnée sur la fig. 15. Les points de calcul disposés sur la

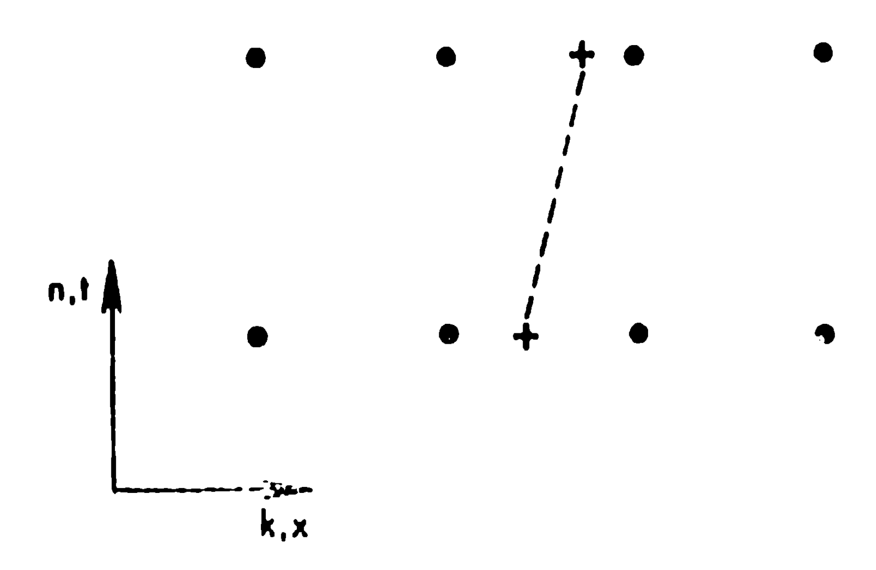


Fig. 15

ligne de discontinuité sont désignés par des croix. On y calcule deux valeurs, de gauche u^- et de droite u^+.

2. Pour le calcul des solutions discontinues de l'équation

$$\frac{\partial U}{\partial t}+\frac{\partial F(U)}{\partial x}=0$$

avec les conditions suivantes sur la ligne de discontinuité $x = X(t)$

$$(U^+ - U^-)\,X' = F(U^+) - F(U^-),$$
$$F'_U(U^-) > F'_U(U^+),$$

où F est une fonction donnée U, appliquer les différentes méthodes possibles d'introduction de la viscosité artificielle. En plus de

$$\frac{\partial U}{\partial t}+\frac{\partial F(U)}{\partial x}+\varepsilon^2\frac{\partial}{\partial x}\left(\frac{\partial U}{\partial x}\right)^2=0$$

envisager l'équation perturbée du type

$$\frac{\partial U}{\partial t}+\frac{\partial F(U)}{\partial x}=\varepsilon\frac{\partial^2 U}{\partial x^2}.$$

Dans les deux cas pour la fonction $U_\varepsilon(x - \omega t)$ étudier les conséquences de l'introduction de la viscosité pour la solu-

tion discontinue. Etudier séparément les mêmes questions pour $F(U)$ linéaire, $F = aU$.

3. Trouver le schéma aux différences pour l'équation (207). Etudier l'approximation et la stabilité.

§ 11. PROBLÈMES MULTIDIMENSIONNELS

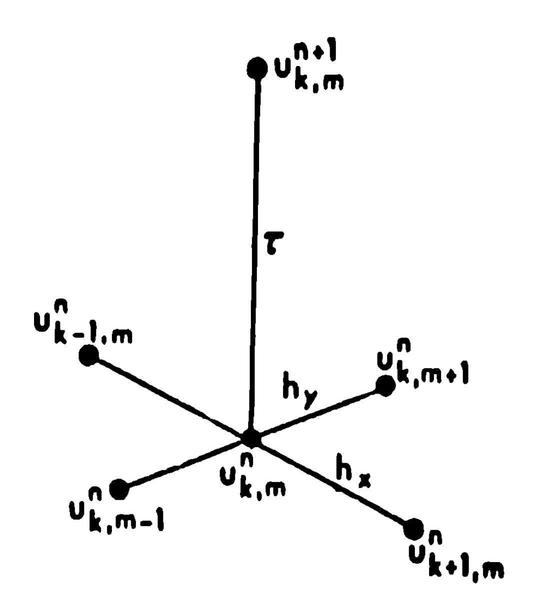
Lorsque l'on passe des équations différentielles ordinaires aux équations aux dérivées partielles, c'est-à-dire lorsque l'on passe d'une variable indépendante à deux, comme nous l'avons déjà vu, non seulement des difficultés quantitatives apparaissent, mais également de nouveaux problèmes importants. Les problèmes essentiels ont été étudiés dans les paragraphes précédents. Le passage aux problèmes à trois et plus variables indépendantes n'est pas non plus trivial. Cependant presque toutes les méthodes envisagées d'élaboration et d'étude des équations aux différences peuvent être simplement et naturellement généralisées à ce cas. Les problèmes à deux variables indépendantes t, x sont à cet effet un bon modèle de problème multidimensionnel.

Nous allons brièvement nous arrêter sur les questions essentielles apparaissant dans les problèmes où les variables indépendantes sont le temps t et deux coordonnées spatiales x, y. Ces problèmes sont appelés *bidimensionnels*. Pour la démonstration nous allons prendre l'équation de la conductibilité thermique, c'est-à-dire le problème

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}; \quad U(0, x, y) = U_0(x, y). \quad (213)$$

Dans le cas bidimensionnel le réseau de calcul simple se composera de points de coordonnées

$$t^n = n\tau, \quad x_k = kh_x, \quad y_m = mh_y,$$



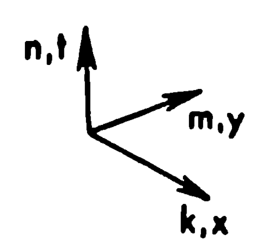


Fig. 16

il est, par conséquent, déterminé par trois paramètres τ, h_x, h_y qui sont les pas du réseau. Désignons par $u^n_{k,\,m}$ les valeurs de la fonction discrète correspondant à ces points.

Toutes les méthodes énumérées au § 7 d'élaboration des formules de calcul peuvent directement être généralisées au cas multidimensionnel.

En particulier, pour le problème (213) lorsque l'on utilise la maille de calcul représentée fig. 16, n'importe laquelle de ces méthodes donne l'équation aux différences

$$\left.\begin{aligned}&\frac{u^{n+1}_{k,\,m}-u^n_{k,\,m}}{\tau}=\\&=\frac{u^n_{k+1,\,m}-2u^n_{k,\,m}+u^n_{k-1,\,m}}{h_x^2}+\frac{u^n_{k,\,m+1}-2u^n_{k,\,m}+u^n_{k,\,m-1}}{h_y^2},\\&u^0_{k,\,m}=U_0(x_k,\;y_m),\end{aligned}\right\}\quad(214)$$

qui est une généralisation directe du cas unidimensionnel (178).

Le schéma de principe de l'étude de la convergence ramenant ce problème à l'approximation et à la stabilité a été exposé dans le § 5 sous une forme très générale, de telle sorte que maintenant le problème bidimensionnel s'en trouve être un cas particulier. Il suffit de remplacer les deux paramètres τ, h et les deux arguments de la fonction discrète n, k par les trois paramètres τ, h_x, h_y et les trois arguments n, k, m.

Pour la vérification de l'approximation du problème aux différences et du problème différentiel on procède d'une manière analogue. Ainsi pour les problèmes (213), (214) supposant la solution exacte $U(t, x, y)$ lisse, on peut écrire

$$U_{k,m}^{n+1} = U_{k,m}^{n} + \tau \left(\frac{\partial U}{\partial t}\right)_{k,m}^{n} + O(\tau^2),$$

$$U_{k\pm 1,m}^{n} = U_{k,m}^{n} \pm h_x \left(\frac{\partial U}{\partial x}\right)_{k,m}^{n} + \frac{1}{2} h_x^2 \left(\frac{\partial^2 U}{\partial x^2}\right)_{k,m}^{n} \pm \frac{1}{6} h_x^3 \left(\frac{\partial^3 U}{\partial x^3}\right)_{k,m}^{n} + O(h_x^4),$$

$$U_{k,m\pm 1}^{n} = U_{k,m}^{n} \pm h_y \left(\frac{\partial U}{\partial y}\right)_{k,m}^{n} + \frac{1}{2} h_y^2 \left(\frac{\partial^2 U}{\partial y^2}\right)_{k,m}^{n} \pm \frac{1}{6} h_y^3 \left(\frac{\partial^3 U}{\partial y^3}\right)_{k,m}^{n} + O(h_y^4).$$

En substituant ces expressions dans (214) on trouve facilement qu'elle est satisfaite à $O(\tau, h_x^2, h_y^2)$ près, c'est-à-dire qu'il y a approximation.

En ce qui concerne l'étude de la stabilité, en fait, nous n'avons étudié (au § 6) qu'une méthode générale, à savoir le critère spectral de stabilité des équations aux différences linéaires de structure stratifiée du type

$$u^{n+1} = Ru^n + \tau f^n.$$

Si par fonction sur la couche u^n on entend maintenant $u^n_{k,\,m}$, c'est-à-dire la fonction discrète de deux indices k, m, tous les raisonnements faits au début du § 6 restent vrais et la stabilité sera une borne pour les normes des puissances de l'opérateur R. Pour les opérateurs R correspondant à la formule

$$(Ru)_{k,\,m}=\sum_{p,\,q}\alpha_{p,\,q}u_{k+p,\,m+q}, \tag{215}$$

qui est une généralisation de (123), on peut trouver l'estimation de ces normes à l'aide du rayon spectral des opérateurs. Ainsi il est facile de voir que les fonctions discrètes

$$u_{k,\,m}=u_{0,\,0}e^{i(k\varphi+m\psi)} \tag{216}$$

pour φ, ψ quelconques sont les fonctions propres de l'opérateur R (215) et

$$\lambda=\sum_{p,\,q}\alpha_{p,\,q}e^{i(p\varphi+q\psi)},$$

les valeurs propres correspondantes. L'inégalité $|\lambda(\varphi,\psi)|\leqslant 1$ donne la condition nécessaire de stabilité. Ainsi dans le cas bidimensionnel, le critère spectral de stabilité conserve son efficacité.

Appliquée au problème (214) cette procédure donne la condition de stabilité suivante

$$\frac{\tau}{h_x^2}+\frac{\tau}{h_y^2}\leqslant\frac{1}{2}.$$

L'approximation utilisant les équations aux différences implicites donne alors des schémas stables quelles que soient les relations entre les pas des réseaux. Pour le problème (213) on aura de toute évidence le schéma suivant (fig. 17)

$$\frac{u^{n+1}_{k,\,m}-u^n_{k,\,m}}{\tau}=\frac{u^{n+1}_{k+1,\,m}-2u^{n+1}_{k,\,m}+u^{n+1}_{k-1,\,m}}{h_x^2}+$$
$$+\frac{u^{n+1}_{k,\,m+1}-2u^{n+1}_{k,\,m}+u^{n+1}_{k,\,m-1}}{h_y^2}. \tag{217}$$

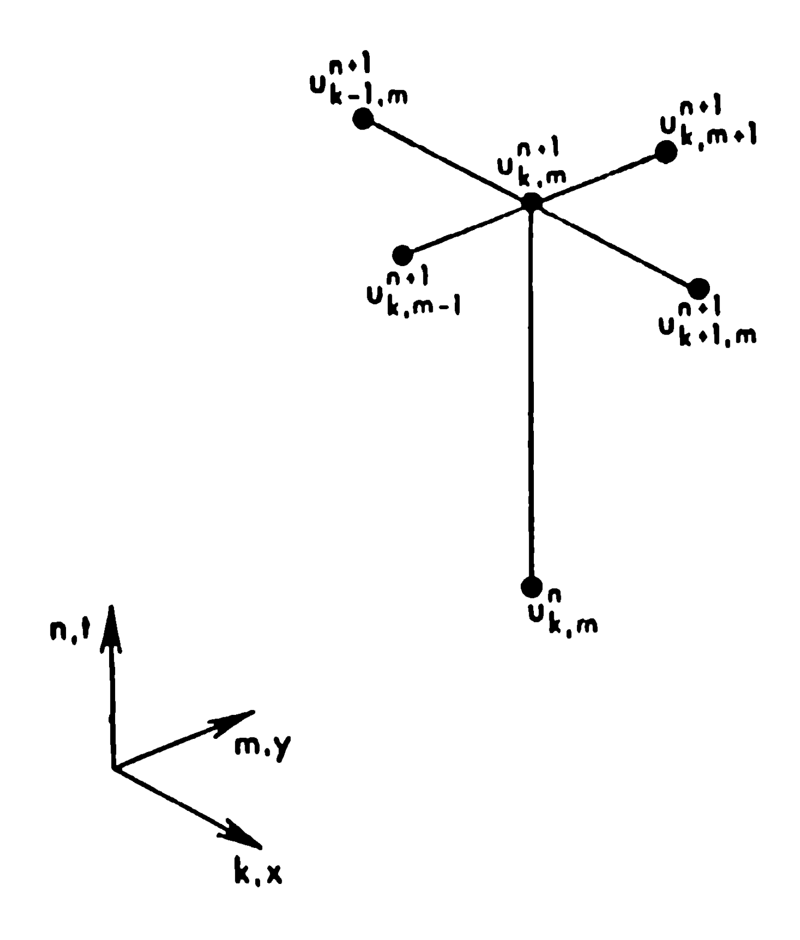


Fig. 17

Sa stabilité peut être vérifiée directement par l'estimation de u^{n+1}, comme dans le § 8. Si pour u^n on prend une fonction du type (216), il est facile de voir que $u^{n+1} = \lambda u^n$, avec

$$\lambda = \frac{1}{1 + \frac{4\tau}{h_x^2} \sin^2 \frac{\varphi}{2} + \frac{4\tau}{h_y^2} \sin^2 \frac{\psi}{2}},$$

et pour τ, h_x, h_y quelconques on a $|\lambda| \leqslant 1$.

Ainsi les questions essentielles liées à l'élaboration et à l'étude des équations aux différences ne changent pas en principe lorsque l'on augmente le nombre de dimensions. Néanmoins l'augmentation du volume du problème et la complication des algorithmes de calcul peuvent donner lieu à de nouveaux problèmes.

Nous allons nous arrêter à cet effet sur les méthodes de résolution des systèmes d'équations aux différences apparais-

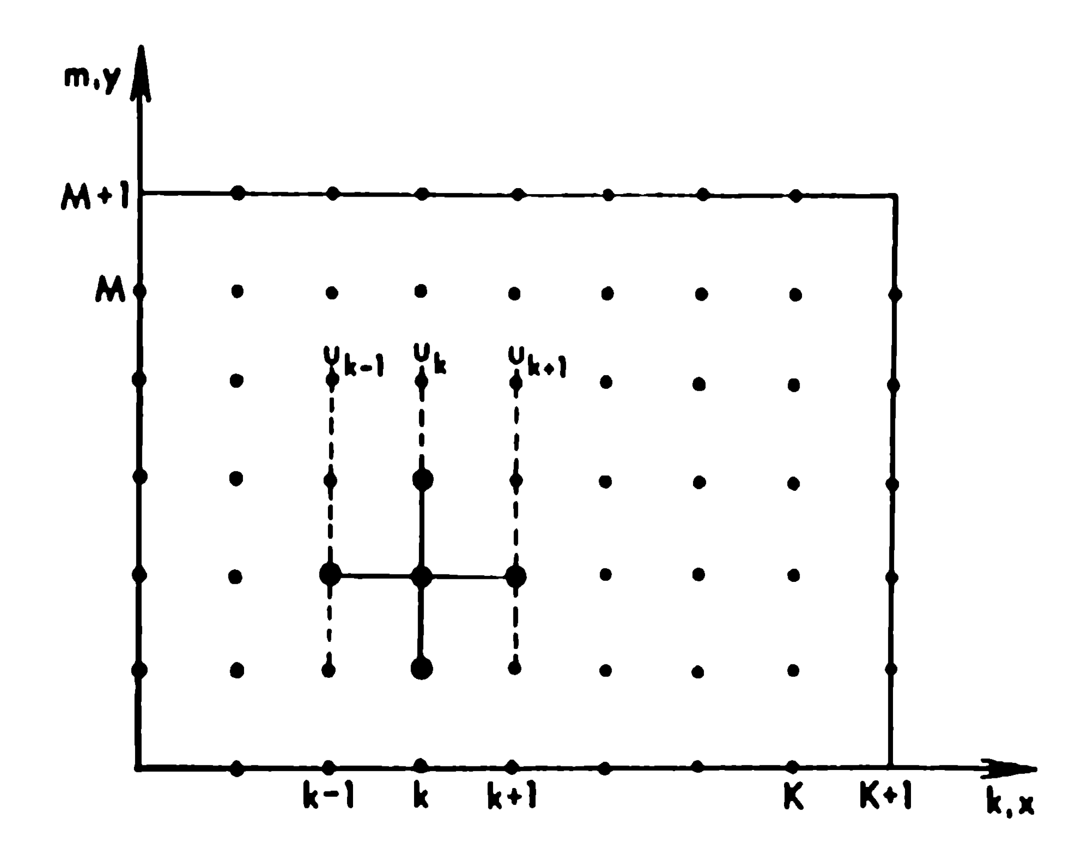


Fig. 18

sant lorsque l'on utilise les schémas implicites. Considérons les équations (217) dans le rectangle (fig. 18)

$$1 \leqslant k \leqslant K, \qquad 1 \leqslant m \leqslant M, \tag{218}$$

en se donnant les valeurs de u^{n+1} sur ses bords. Pour simplifier on pose $h_x = h_y = h$, on introduit la désignation $r = \tau/h^2$, on omet l'indice $n+1$ auprès des inconnues $u_{k,m}^{n+1}$ et on peut alors écrire ces équations comme suit

$$-ru_{k,m+1} - ru_{k-1,m} + (1+4r)\,u_{k,m} - ru_{k+1,m} - \\ - ru_{k,m-1} = u_{k,m}^n. \tag{219}$$

Les valeurs aux limites de $u_{k,m}$ sont données comme suit

$$\left.\begin{array}{c} u_{k,M+1} = \delta_k, \\ u_{0,m} = \alpha_m, \quad u_{K+1,m} = \beta_m, \\ u_{k,0} = \gamma_k. \end{array}\right\} \tag{220}$$

Les indices k, m prennent l'ensemble de valeurs (218). Ainsi le nombre d'équations et le nombre d'inconnues sont égaux à KM, à chaque point intérieur du rectangle correspond une équation propre (219).

Dans le cas unidimensionnel, pour la solution d'un système analogue d'équations on a pu utiliser la méthode efficace d'exclusion, à savoir la méthode du balayage (§ 9), car nous avions un système très simple, où chaque équation ne contenant que trois inconnues de numéros successifs, correspondant à leur disposition naturelle. Ici il n'en est rien, néanmoins on peut élaborer une méthode de résolution du système d'équations (219) généralisant la méthode du balayage qui nous a si bien servi dans le cas unidimensionnel.

Nous allons considérer l'ensemble des valeurs $u_{k,\,1}$, $u_{k,\,2}$, . . ., $u_{k,\,M}$ pour k donné, comme les composantes du vecteur u_k de dimension M (fig. 18). Choisissons parmi toutes les équations du système (219) celles correspondant à cette valeur de k. Il est évident qu'elles relieront les composantes de trois vecteurs seulement u_{k-1}, u_k, u_{k+1}. Ecrivons ces M équations sous la forme d'une seule équation vectorielle

$$-Au_{k-1}+Bu_k-Cu_{k+1}=d_k, \tag{221}$$

où A, B, C sont des matrices carrées, et d_k, un vecteur d'ordre M.

Il est évident que $A = C = cI$ (I étant la matrice unité), et

$$B=\begin{pmatrix} 1+4r & -r & 0 & & & 0 \\ -r & 1+4r & -r & & & \\ 0 & -r & 1+4r & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & & \ddots & 1+4r & -r \\ 0 & & & & -r & 1+4r \end{pmatrix},$$

$$d_k = \begin{pmatrix} u^n_{k,1} + r\gamma_k \\ u^n_{k,2} \\ u^n_{k,3} \\ \vdots \\ u^n_{k,M-1} \\ u^n_{k,M} + r\delta_k \end{pmatrix}.$$

Comme le système (219) s'est réduit à K équations (221) reliant des vecteurs u_{k-1}, u_k, u_{k+1} par trois, on peut utiliser la méthode du balayage, tenant compte évidemment du caractère vectoriel des équations (221).

Supposons qu'entre u_{k-1} et u_k on ait la relation

$$u_{k-1} = L_k u_k + M_k, \tag{222}$$

où L_k est une matrice carrée, et M_k un vecteur du même ordre que u_k. En substituant (222) dans (221), on exclut u_{k-1} et l'on obtient donc une relation entre la paire suivante de vecteurs

$$(B - AL_k)\, u_k - Cu_{k+1} = AM_k + d_k.$$

En résolvant cette dernière par rapport à u_k, c'est-à-dire en multipliant à gauche par la matrice inverse de $B - AL_k$, on obtient

$$u_k = (B - AL_k)^{-1}(Cu_{k+1} + AM_k + d_k).$$

Pour écrire cette dernière relation sous la forme (222), on pose

$$\left.\begin{aligned} L_{k+1} &= (B - AL_k)^{-1} C, \\ M_{k+1} &= (B - AL_k)^{-1}(AM_k + d_k). \end{aligned}\right\} \tag{223}$$

La méthode de résolution est maintenant claire. La condition aux limites pour $k = 0$ détermine $L_1 = 0$, $M_1 = \alpha$. A l'aide des formules (223) on trouve successivement tous les L_k, M_k. Comme on connaît u_{K+1}, $u_{K+1} = \beta$, à partir de L_k, M_k en utilisant la formule (222) on obtient toutes

les solutions u_k du système. Les coefficients L_k étant des matrices, la méthode exposée est appelée *méthode du balayage matriciel*.

Apparemment elle est analogue à la méthode unidimensionnelle du balayage. Cependant à l'opposé de celle-ci elle est très rarement utilisée, ce qui est dû à ce qu'elle conduit à des calculs compliqués. Lors de chaque cycle du processus de calcul il faut prendre l'inverse d'une matrice d'ordre élevé $\sim 1/h$ (et fixer dans la mémoire $\sim 1/h$ de ces matrices). C'est pourquoi souvent l'utilisation des méthodes itératives de résolution des systèmes (et même l'utilisation des schémas explicites) se trouve être plus efficace.

Lors de l'élaboration d'un algorithme plus économique pour le type de problèmes envisagés on procède différemment. Considérons le schéma aux différences (fig. 19)

$$\frac{u_{k,m}^{n+1}-u_{k,m}^{n}}{\tau}=\frac{u_{k+1,m}^{n+1}-2u_{k,m}^{n+1}+u_{k-1,m}^{n+1}}{h_x^2}+$$
$$+\frac{u_{k,m+1}^{n}-2u_{k,m}^{n}+u_{k,m-1}^{n}}{h_y^2}, \qquad (224)$$

qui est intermédiaire entre (214), (217) donnant également une approximation de l'équation différentielle (213). En substituant dans (224) les fonctions discrètes du type (216), on trouve $u^{n+1}=\lambda u^n$, où

$$\lambda=\frac{1-\frac{4\tau}{h_y^2}\sin^2\frac{\psi}{2}}{1+\frac{4\tau}{h_x^2}\sin^2\frac{\varphi}{2}},$$

et par conséquent le schéma aux différences (224) ne peut être stable que si

$$\frac{\tau}{h_y^2}\leqslant\frac{1}{2}, \qquad (225)$$

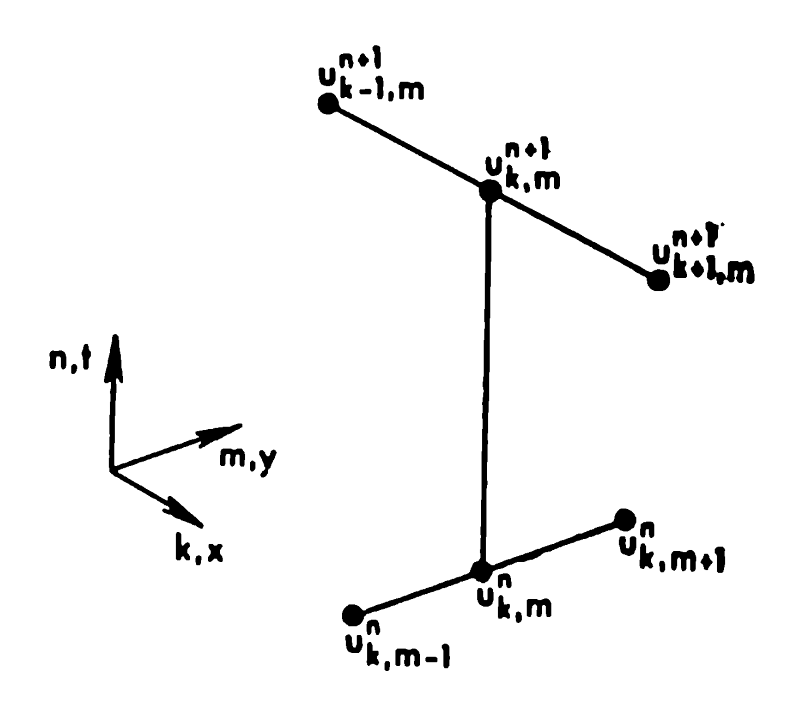


Fig. 19

ce qui n'est pas étonnant car ce schéma est explicite par rapport à l'une des variables, à savoir y. On utilise de tels schémas si le problème impose que les directions x et y ne soient pas équivalentes. Par exemple si la solution dépend faiblement de y, on peut prendre h_y bien plus grand que h_x, la condition (225) n'est alors pas difficile à remplir.

Remarquons le fait que d'un côté le schéma (224) n'impose aucune restriction à la relation existant entre τ et h_x et d'un autre côté, l'ensemble des équations (224) pour chaque m donné forme un système pouvant être résolu par la méthode simple du balayage.

On peut dire ainsi que le problème de l'élaboration d'un algorithme efficace est à moitié résolu. Il ne reste plus que la condition imposée à τ et h_y (225). Le nombre d'opérations arithmétiques nécessaires pour obtenir la solution se trouve être proportionnel au nombre de points de calcul.

Changeons de rôle les directions x et y, c'est-à-dire envisageons un schéma explicite par rapport à x et implicite par rapport à y. On obtient un schéma permettant de résoudre la seconde moitié du problème. Les schémas s'excluent

mutuellement, cependant nous allons essayer de les utiliser successivement, l'un pour les pas pairs et l'autre pour ceux impairs par rapport à t. Comme le cycle élémentaire sera dans ce cas une paire de pas, il est plus commode d'appeler un pas dans le temps tout le cycle, d'utiliser chaque pas pour un déplacement de $\tau/2$. La solution obtenue sur la première moitié du pas est alors une certaine solution intermédiaire $\tilde{u}$.

Ceci donne le schéma aux différences suivant

$$\frac{\tilde{u}_{k,m}-u^n_{k,m}}{\tau/2}=\frac{\tilde{u}_{k+1,m}-2\tilde{u}_{k,m}+\tilde{u}_{k-1,m}}{h_x^2}+\frac{u^n_{k,m+1}-2u^n_{k,m}+u^n_{k,m-1}}{h_y^2}, \qquad (226)$$

$$\frac{u^{n+1}_{k,m}-\tilde{u}_{k,m}}{\tau/2}=\frac{\tilde{u}_{k+1,m}-2\tilde{u}_{k,m}+\tilde{u}_{k-1,m}}{h_x^2}+\frac{u^{n+1}_{k,m+1}-2u^{n+1}_{k,m}+u^{n+1}_{k,m-1}}{h_y^2}. \qquad (227)$$

Comme précédemment, c'est une approximation de l'équation (213). Nous allons étudier la stabilité de ce schéma.

En utilisant pour u^n une fonction du type (216), on obtient à partir de (226) $\tilde{u}=\tilde{\lambda}u^n$, où

$$\tilde{\lambda}=\frac{1-\frac{2\tau}{h_y^2}\sin^2\frac{\psi}{2}}{1+\frac{2\tau}{h_x^2}\sin^2\frac{\varphi}{2}},$$

et à partir de (227) $u^{n+1}=\tilde{\tilde{\lambda}}\tilde{u}$, où

$$\tilde{\tilde{\lambda}}=\frac{1-\frac{2\tau}{h_x^2}\sin^2\frac{\varphi}{2}}{1+\frac{2\tau}{h_y^2}\sin^2\frac{\psi}{2}}.$$

Seul le produit $\tilde{\lambda}\tilde{\tilde{\lambda}} = \lambda$ nous intéresse car c'est lui qui correspond à un pas entier dans le temps. Il est facile de voir que λ est le produit de deux facteurs

$$\lambda = \frac{1 - \frac{2\tau}{h_x^2} \sin^2 \frac{\varphi}{2}}{1 + \frac{2\tau}{h_x^2} \sin^2 \frac{\varphi}{2}} \cdot \frac{1 - \frac{2\tau}{h_y^2} \sin^2 \frac{\psi}{2}}{1 + \frac{2\tau}{h_y^2} \sin^2 \frac{\psi}{2}}, \qquad (228)$$

chacun de ces facteurs n'étant pas supérieur en module à l'unité pour τ, h_x, h_y, φ, ψ quelconques.

Ainsi le schéma aux différences (226), (227), tout comme le schéma implicite (217), est stable quelles que soient les relations existant entre les pas de réseau et, à la différence de (217), est économique du point de vue du nombre des opérations. En effet, le processus de résolution des équations (226), (227) revient tout d'abord à la resolution du système (226) pour tout m donné, à la recherche de $\tilde{u}$ et ensuite à la résolution du système (227) pour chaque k donné, ce qui donne u^{n+1}. Les deux peuvent être résolus par la méthode habituelle du balayage. Les méthodes aux différences de ce type sont appelées *méthodes des directions alternées* ou *méthodes des pas fractionnaires.*

Il est facile de comprendre pourquoi la méthode du balayage matriciel se trouve être moins efficace que celle qui vient d'être exposée: elle est trop universelle. En effet, on sait que les avantages de la méthode habituelle du balayage sont dus à ce qu'on tient compte avec une grande précision de l'influence mutuelle des solutions dans les différents points. Considérons la méthode du balayage matriciel de ce point de vue. La relation entre les vecteurs u_{k-1}, u_k écrite sous la forme (222) fait apparaître formellement les relations existant entre les composantes de ces vecteurs. Il est cependant évident que l'influence mutuelle des différentes composantes diminue rapidement au fur et à mesure de l'éloignement réciproque des points de calcul corres-

pondants lors de l'augmentation de la différence entre les numéros m. La méthode du balayage matriciel ne tient pas compte de cette particularité du système d'équations, elle vise une classe bien plus étendue de problèmes et dans le cas présent est loin d'être la meilleure.

Nous avons envisagé seulement un seul des problèmes apparaissant lorsque l'on passe des problèmes unidimensionnels aux problèmes bidimensionnels. Tout comme précédemment nous avons à cet effet utilisé un exemple particulier caractéristique permettant de faire apparaître l'essence du problème. Il est évident qu'en augmentant le nombre des dimensions on voit apparaître de nouveaux problèmes, mais nous n'allons pas nous y arrêter. Ceux-ci sont essentiellement liés aux difficultés d'approximation des domaines multidimensionnels par des réseaux de calcul convenables et à l'obtention d'un algorithme de calcul simple et économique.

Problèmes

1. Trouver et étudier les différents schémas pour la solution de l'équation

$$\frac{\partial U}{\partial t}+a\frac{\partial U}{\partial x}+b\frac{\partial U}{\partial y}=0,$$

en utilisant la maille de calcul représentée fig. 16.

Etudier également le schéma aux différences du type

$$\frac{u_{k,m}^{n+1}-u_{k,m}^{n-1}}{2\tau}+a\frac{u_{k+1,m}^{n}-u_{k-1,m}^{n}}{2h_x}+b\frac{u_{k,m+1}^{n}-u_{k,m-1}^{n}}{2h_y}=0.$$

2. Etudier les méthodes itératives pour la résolution des systèmes d'équations (219), (220). Donner une estimation du nombre d'itérations.

3. Donner une estimation du nombre d'opérations arithmétiques nécessaires à la résolution des problèmes sur un

intervalle de temps fini et dans un domaine limité, en utilisant

a) le schéma explicite (214),

b) le schéma implicite (217) résolu par la méthode du balayage matriciel,

c) le schéma implicite (217) résolu par les méthodes itératives,

d) la méthode des directions alternées (226), (227).

Dans les trois derniers cas poser $\tau \sim h$.

4. Eliminer entre les formules (226), (227) la grandeur intermédiaire $\tilde{u}$. Comparer l'équation aux différences obtenue avec le schéma implicite (217).

5. Montrer que pour la solution du problème (213) on peut utiliser les algorithmes aux différences donnés par les formules

$$\frac{\tilde{u}_{k,m}-u^n_{k,m}}{\tau}=\frac{\tilde{u}_{k+1,m}-2\tilde{u}_{k,m}+\tilde{u}_{k-1,m}}{h_x^2}+\frac{u^n_{k,m+1}-2u^n_{k,m}+u^n_{k,m-1}}{h_y^2},$$

$$\frac{u^{n+1}_{k,m}-\tilde{u}_{k,m}}{\tau}=\frac{u^{n+1}_{k,m+1}-2u^{n+1}_{k,m}+u^{n+1}_{k,m-1}}{h_y^2}-\frac{u^n_{k,m+1}-2u^n_{k,m}+u^n_{k,m-1}}{h_y^2}$$

et

$$\frac{\tilde{u}_{k,m}-u^n_{k,m}}{\tau}=\frac{\tilde{u}_{k+1,m}-2\tilde{u}_{k,m}+\tilde{u}_{k-1,m}}{h_x^2},$$

$$\frac{u^{n+1}_{k,m}-\tilde{u}_{k,m}}{\tau}=\frac{u^{n+1}_{k,m+1}-2u^{n+1}_{k,m}+u^{n+1}_{k,m-1}}{h_y^2}.$$

Les comparer avec le schéma implicite (217) et la méthode des directions alternées (226), (227) (en excluant $\tilde{u}$).

6. Envisager la possibilité de généralisation de toutes les méthodes mentionnées dans le paragraphe présent aux problèmes tridimensionnels correspondants.

§ 12. PROBLÈMES STATIONNAIRES

On utilise ce terme pour les problèmes décrivant des états stationnaires des différents systèmes ne changeant pas dans le temps. A titre d'exemple typique on peut étudier le problème suivant. Soit à trouver la fonction $U(x, y)$ satisfaisant dans un certain domaine limité G du plan x, y à l'équation

$$\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}=f(x, y), \tag{229}$$

prenant sur la limite Γ de ce domaine des valeurs données, à savoir

$$U|_{\Gamma}=g. \tag{230}$$

Il n'est pas difficile de trouver le problème aux différences correspondant. Recouvrons le domaine G par un réseau de calcul, pour plus de simplicité les pas suivant x et y sont pris égaux (fig. 20). Sur ce réseau on remplace l'équation (229) par la relation aux différences

$$\frac{u_{k+1, m}-2u_{k, m}+u_{k-1, n}}{h^2}+\frac{u_{k, m+1}-2u_{k, m}+u_{k, m-1}}{h^2}=f_{k, m}, \tag{231}$$

ayant un sens pour chaque point interne de calcul. On entendra par point interne un point quelconque k, m pour lequel les quatre points voisins $k \pm 1$, $m \pm 1$ utilisés dans (231) sont disposés à l'intérieur du domaine G. Les autres points de calcul k, m appartenant à G seront dits limites, leur

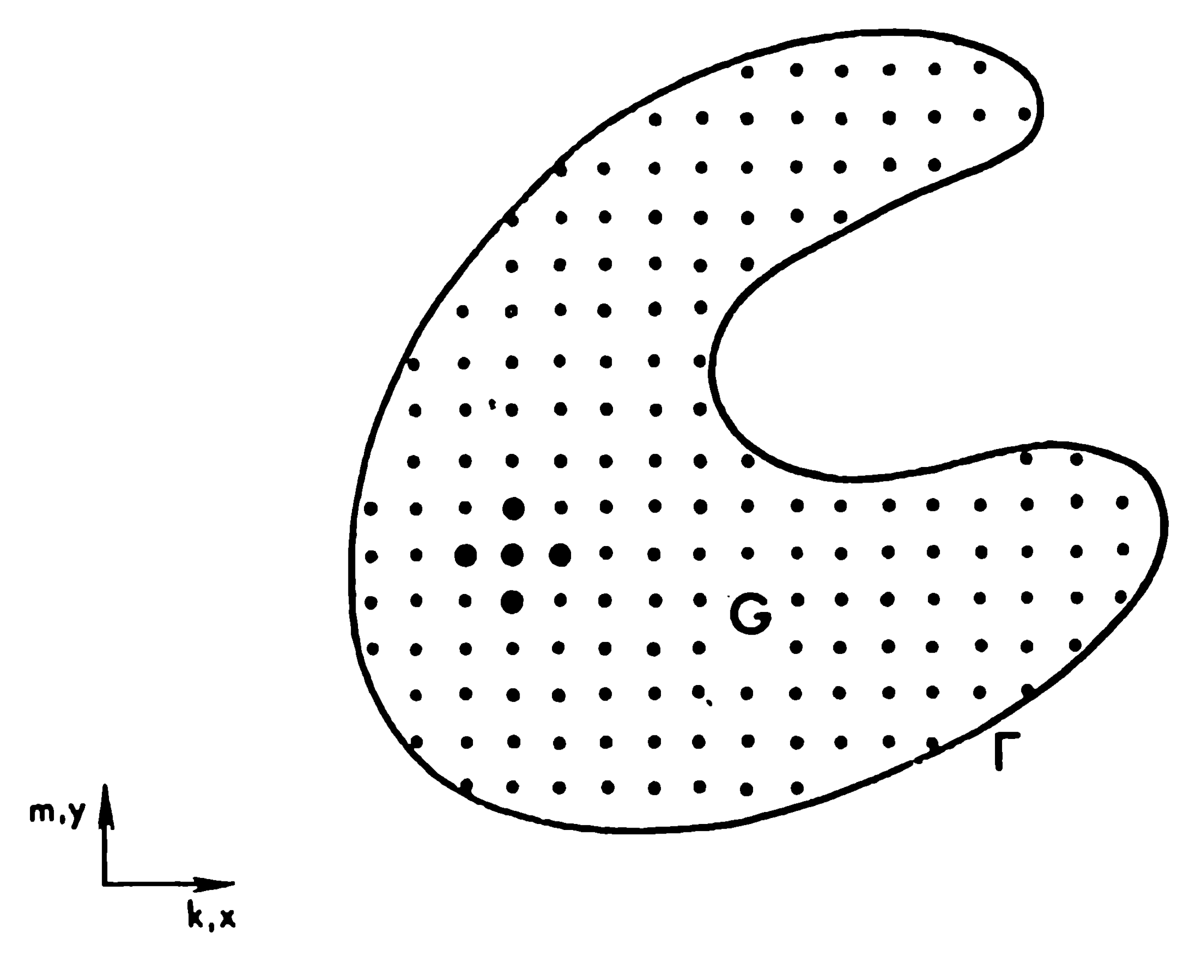


Fig. 20

ensemble étant désigné par γ. On obtient les valeurs $u_{k,\ m}$ sur γ simplement par translation de la valeur g à partir du point voisin de la frontière Γ

$$u|_\gamma = g|_\Gamma. \qquad (232)$$

Le problème aux différences (231), (232) appartient à la classe envisagée au § 5. Son étude revient à la vérification de l'approximation et de la stabilité. La première est évidente car à la limite, pour $h \to 0$, (231) devient (229) et (232) devient (230) (la distance entre γ et Γ étant de l'ordre de h).

Nous allons démontrer la stabilité du problème aux différences établi, c'est-à-dire la coïncidence des ordres de grandeurs de la solution et des seconds membres de (231), (232) quel que soit h. A cet effet on utilise l'artifice suivant.

Supposons que la solution u du problème existe. Considérons les deux fonctions auxiliaires v_+ et v_-

$$v_{\pm} = \pm u + \alpha (x^2 + y^2) + \beta, \tag{233}$$

où α, β sont pour le moment des constantes arbitraires. Désignons par Du le premier membre de (231). Substituons (233) dans (231), il vient

$$Dv_{\pm} = \pm f + 4\alpha,$$

car $Du = f$, $D(x^2 + y^2) = 4$. $D\beta = 0$.

On choisit α de telle sorte que partout dans le domaine G on ait l'inégalité

$$Dv_{\pm} \geqslant 0. \tag{234}$$

Il est évident que pour cela il suffit de poser

$$\alpha = \frac{1}{4} \max_{k,\, m} |f_{k,\, m}|. \tag{235}$$

Puis on détermine β de telle sorte que

$$v_{\pm}|_{\gamma} \leqslant 0. \tag{236}$$

En vertu de (232), (233) ceci aura lieu pour

$$\beta = -\max_{\Gamma} |g| - \alpha \max_{G} (x^2 + y^2). \tag{237}$$

Supposons que max $v_{\pm}$ corresponde à un certain point interne k, m. Dans ce cas au moins dans un des points voisins la valeur v_+ est inférieure à la valeur maximale, et comme on peut facilement le voir à partir de (231), $(Dv_{\pm})_{k,\, m} < 0$. Ceci se trouve en contradiction avec (234) et par conséquent max $v_{\pm}$ ne peut se trouver que sur la frontière γ. Mais, vu (236), $v_{\pm}$ y est négative, elle est donc négative partout, ainsi

$$\pm u_{k,\, m} + \alpha (x^2 + y^2)_{k,\, m} + \beta \leqslant 0,$$

c'est-à-dire en vertu de (235), (237), on a

$$\max_{k, m} |u_{k, m}| \leqslant \max_{\Gamma} |g| + \frac{1}{4} \max_{k, m} |f_{k, m}| \max_{G} (x^2 + y^2). \quad (238)$$

L'inégalité obtenue signifie que le problème (231), (232) est stable.

Nous avons démontré également que la solution existe et qu'elle est unique. En effet comme chaque solution doit satisfaire à (238), alors pour $g = f = 0$ seule la solution triviale $u = 0$ est possible. Par conséquent la solution du système d'équations linéaires non homogènes (231), (232) existe et est unique.

Abordons maintenant les méthodes de résolution du système d'équations (231), (232). Du point de vue historique les méthodes itératives sont apparues les premières. Nous allons décrire la plus simple d'entre elles.

Résolvons chacune des équations (231) par rapport à la valeur $u_{k, m}$ au point central de la maille de calcul:

$$u_{k, m} = \frac{1}{4} (u_{k-1, m} + u_{k+1, m} + u_{k, m-1} + u_{k, m+1} - h^2 f_{k, m}) \quad (239)$$

et utilisons cette formule pour les itérations. Le processus de calcul est très simple: sur chaque ν-ième itération on calcule la moyenne arithmétique des valeurs $u^{(\nu)}_{k\pm 1, m\pm 1}$ aux points entourant le point central donné et l'on obtient l'approximation suivante $u^{(\nu+1)}_{k, m}$.

Etudions la convergence du processus. Posons

$$u^{(\nu)}_{k, m} = u_{k, m} + \delta^{(\nu)}_{k, m},$$

où $u_{k, m}$ est la solution exacte du système (231), (232). Il est dans ce cas évident que l'erreur $\delta_{k, m}$ sera déterminée par le processus itératif suivant

$$\delta^{(\nu+1)}_{k, m} = \frac{1}{4} (\delta^{(\nu)}_{k+1, m} + \delta^{(\nu)}_{k-1, m} + \delta^{(\nu)}_{k, m+1} + \delta^{(\nu)}_{k, m-1}), \quad \delta^{(\nu+1)}|_{\gamma} = 0. \quad (240)$$

Désignons par $\bar{\delta}$ le maximum du module de $\delta_{k,m}^{(0)}$ nous allons raisonner comme suit. Comme $\delta_{k,m}^{(1)}$ est la moyenne arithmétique des quatre valeurs $\delta_{k\pm1,m\pm1}^{(0)}$, $|\delta_{k,m}^{(1)}|$ ne surpasse pas $\bar{\delta}$, ce qui est vrai pour tous les points sur toutes les itérations. Mais pour les points voisins des points limites on peut donner une estimation plus précise. A savoir, si ne serait-ce qu'un des points voisins de k, m est un point limite, où $\delta = 0$, en ce point k, m on a

$$|\delta_{k,m}^{(1)}| \leqslant \frac{0+3\bar{\delta}}{4} = \frac{3}{4}\bar{\delta}.$$

Cette dernière inégalité est vraie pour toute la couche limitrophe de points. Passons à la seconde itération, où l'influence de la frontière s'étend encore à une couche de points où

$$|\delta_{k,m}^{(2)}| \leqslant \frac{\frac{3}{4}\bar{\delta}+3\bar{\delta}}{4} = \frac{15}{16}\bar{\delta}.$$

Cette estimation est a fortiori vraie pour la première couche limitrophe de points. En continuant le raisonnement, on va se déplacer lors de chaque itération à l'intérieur du domaine G, pour les couches de points traversées on obtient alors l'estimation

$$|\delta_{h,m}^{(\nu)}| \leqslant \left(1-\frac{1}{4^{\nu}}\right)\bar{\delta}.$$

Enfin lors d'une certaine itération d'ordre n, $n \sim 1/h$, tous les points de calcul se trouvent épuisés. Ceci signifie que lors de n itérations l'erreur δ a diminué au moins de $(1-4^{-n})$ fois. Lors des n itérations suivantes, elle diminue encore du même nombre de fois, etc. Nous avons démontré que pour $\nu \to \infty$ l'erreur $\delta^{(\nu)} \to 0$, c'est-à-dire que le processus itératif converge.

Pour n itérations l'erreur diminue de $(1-4^{-n})$ fois, par conséquent pour une itération elle diminue en moyenne de

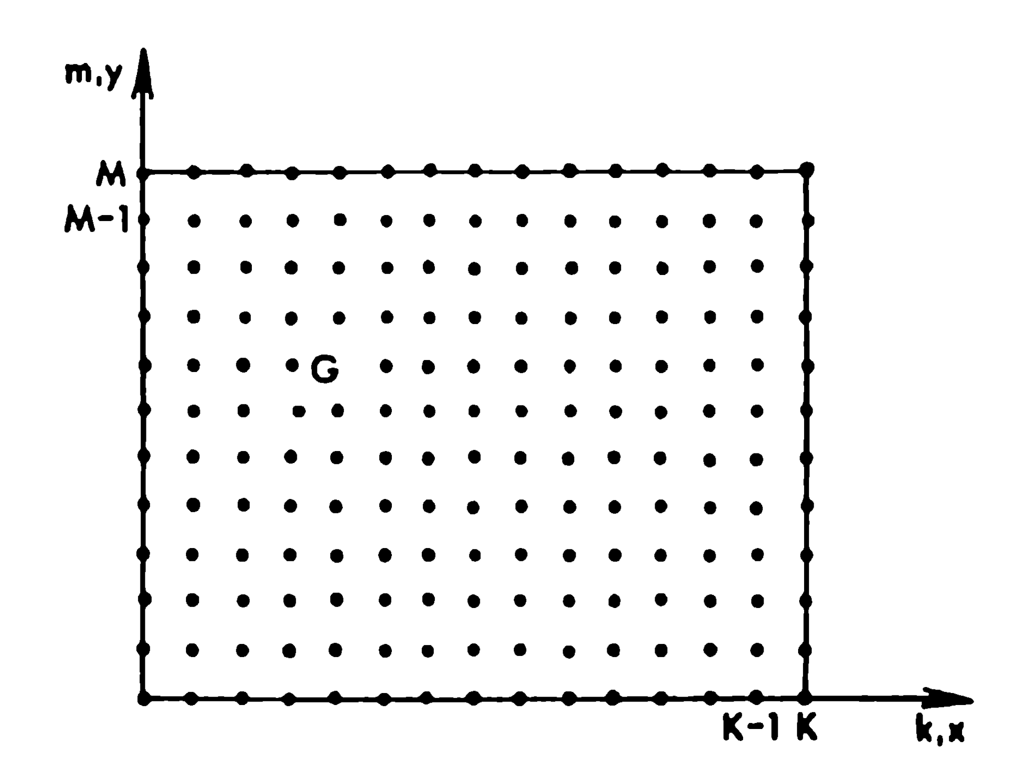


Fig. 21

$(1-4^{-n})^{1/n}$ fois, ou (comme $n \sim 1/h$) de $(1-4^{-1/h})^h \sim$ $\sim 1 - h4^{-1/h}$ fois. Généralement on caractérise la *vitesse de convergence* par la grandeur de la décroissance relative de l'erreur pour une itération, c'est-à-dire par le rapport $(\delta^{(\nu)} - \delta^{(\nu+1)})\ \delta^{(\nu)} = \varkappa$. Dans le cas présent pour cette grandeur on obtient l'estimation suivante

$$\varkappa \sim h4^{-1/h}. \tag{241}$$

En fait, dans de nombreux cas cette estimation est trop surhaussée, le processus itératif converge plus rapidement. Ainsi, si le domaine G est un rectangle (fig. 21), il est facile d'obtenir une estimation plus précise. A cet effet remarquons que la formule (240) décrivant l'évolution de l'erreur d'une itération à l'autre peut être interprétée comme une formule d'un problème aux différences

$$\delta^{(\nu+1)} = R\delta^{(\nu)}$$

d'une classe que nous connaissons déjà (§ 6). La seule différence est que ν est maintenant le numéro de l'itération et non le numéro de la couche temporelle. C'est pourquoi pour l'étude de l'évolution de l'erreur $\delta^{(\nu)}$ on peut appliquer le critère spectral. Pour le moment on ne tient pas compte des conditions aux limites $\delta|_\gamma = 0$ et l'on pose

$$\delta^{(\nu)}_{k,\,m} = \delta^{(\nu)}_{0,\,0} e^{i(k\varphi + m\psi)}. \tag{242}$$

Comme toujours, $\delta^{(\nu+1)}$ est égal à $\lambda\delta^{(\nu)}$ et dans le cas présent, on obtient facilement que

$$\lambda = \frac{\cos\varphi + \cos\psi}{2}, \tag{243}$$

c'est-à-dire $|\lambda| \leqslant 1$. Cette estimation nous assurait la stabilité des problèmes d'évolution, mais maintenant elle devient insuffisante. Seule l'inégalité rigoureuse $|\lambda| < 1$ nous convient, car ce n'est que dans ce cas que la convergence est garantie: $\delta^{(\nu)} \to 0$ pour $\gamma \to \infty$. Il est facile de voir que $|\lambda| = 1$ correspond aux fonctions (242) qui ne satisfont pas aux conditions aux limites $\delta|_\gamma = 0$ (celles-ci s'obtiennent pour $\varphi = \psi = 0$ ou $\varphi = \psi = \pi$), donc l'estimation $|\lambda|$ peut être précisée.

Nous ne laisserons que les combinaisons des fonctions (242) qui s'annulent sur les limites de notre domaine rectangulaire G (fig. 21), c'est-à-dire satisfont à la condition

$$\delta_{0,\,m} = \delta_{K,\,m} = 0, \qquad m = 0,\ 1,\ \ldots,\ M,$$
$$\delta_{k,\,0} = \delta_{k,\,M} = 0, \qquad k = 0,\ 1,\ \ldots,\ K.$$

Comme les exponentielles figurant dans (242) s'expriment en fonction de $\sin k\varphi$, $\cos k\varphi$, $\sin m\psi$, $\cos m\psi$, les fonctions nous intéressant sont des combinaisons de ces dernières. Pour satisfaire à la condition limite de gauche $\delta_{0,\,m} = 0$, ces fonctions doivent contenir le facteur $\sin k\psi$. Si l'on exige que soit vérifiée la condition limite sur l'extrémité droite, pour $k = K$, on obtient l'égalité $K\varphi = 0$. Ceci

n'est possible que pour $K\varphi = p\pi$, où p est entier. Ainsi nous devons considérer seulement un ensemble discret φ et même fini

$$\varphi_p = p\frac{\pi}{K}, \qquad p = 1, 2, \ldots, K-1, \tag{244}$$

car pour les autres p on obtient les mêmes fonctions discrètes $\sin k\varphi_p$, et pour $p = 0$ ou $p = K$, une fonction nulle sur le réseau.

Pour des raisons analogues nos fonctions $\delta_{k,m}$ doivent contenir le facteur $\sin m\psi_q$ avec

$$\psi_q = q\frac{\pi}{M}, \qquad q = 1, 2, \ldots, M-1. \tag{245}$$

Ainsi les fonctions discrètes

$$\delta_{k,m} = \sin k\varphi_p \sin m\psi_q \tag{246}$$

pour $\varphi_p\psi_q$ quelconques définis par les égalités (244), (245), satisfont aux conditions aux limites.

Substituons (246) dans (240) au lieu de $\delta_{k,m}^{(\nu)}$. Après des calculs simples on obtient :

$$\delta_{k,m}^{(\nu+1)} = \frac{\cos\varphi_p + \cos\psi_q}{2}\sin k\varphi_p \sin m\psi_q,$$

c'est-à-dire $\delta_{k,m}$ (246) sont des fonctions propres de l'opérateur d'itérations, et les valeurs propres correspondantes sont données par la formule

$$\lambda_{p,q} = \frac{\cos\varphi_p + \cos\psi_q}{2}, \tag{247}$$

coïncidant avec (243). La réserve des fonctions propres est grande (on peut montrer qu'elle est suffisamment grande) et la grandeur $\lambda_{p,q}$ (247) permet de juger de la vitesse réelle de la convergence du processus itératif.

Les plus grandes valeurs de $|\lambda_{p,q}|$ sont atteintes pour les valeurs limites φ_p, ψ_q, c'est-à-dire pour $|\cos\varphi_p| = \cos(\pi/K)$ et $|\cos\psi_q| = \cos(\pi/M)$. Comme Kh et Mh

déterminent les dimensions du domaine G, c'est-à-dire qu'ils sont de l'ordre de l'unité, on peut écrire

$$\max_{p,\,q} |\lambda_{p,\,q}| = \frac{\cos(\pi/K) + \cos(\pi/M)}{2} \sim \cos h \sim 1 - \frac{h^2}{2}\ .$$

Par conséquent la caractéristique introduite ci-dessus de la vitesse de convergence des itérations est

$$\varkappa \sim h^2, \tag{248}$$

ce qui est évidemment mieux que (241) obtenu par une estimation grossière.

Nous avons déjà noté l'analogie existant entre le processus itératif et le problème d'évolution aux différences. En fait, elle va bien plus loin. Chaque problème stationnaire peut être considéré comme un cas particulier d'un problème d'évolution, où seul nous intéresse l'état final, stationnaire, et non le processus de l'établissement. Ceci influe également sur les méthodes de résolution des problèmes stationnaires. Tous ces problèmes sont itératifs, à partir du plus simple: calcul des racines d'une équation, ils utilisent tous l'évolution, même fictive. Les problèmes linéaires font exception, mais si l'on tient compte du fait que même la division est une opération itérative, il devient évident que cette exception confirme la règle générale. Nous avons commencé par écrire l'équation $f(x) = 0$ sous la forme $x = \varphi(x)$ (§ 1). Nous allons terminer en répétant cette procédure pour le problème du présent paragraphe.

Ainsi pour la résolution des problèmes stationnaires on peut appliquer tout l'appareil élaboré pour les problèmes d'évolution. La formule envisagée du processus itératif (239), peut s'écrire sous la forme

$$\frac{u^{(\nu+1)}_{k,\,m} - u^{(\nu)}_{k,\,m}}{h^2/4} = \frac{u^{(\nu)}_{k-1,\,m} - 2u^{(\nu)}_{k,\,m} + u^{(\nu)}_{k+1,\,m}}{h^2} + \\ + \frac{u^{(\nu)}_{k,\,m-1} - 2u^{(\nu)}_{k,\,m} + u^{(\nu)}_{k,\,m+1}}{h^2} - f_{k,\,m}.$$

Si ν est le numéro de la couche dans le temps, et $h^2/4$, le pas τ, on obtient le schéma aux différences (214) pour la résolution de l'équation bidimensionnelle de la conductibilité thermique que nous connaissons déjà. Remarquons que sa condition de stabilité se trouve vérifiée car $\tau/h^2 = 1/4$.

Comme nous l'avons vu, la meilleure méthode de résolution du problème mentionné s'est trouvée être la méthode des directions alternées. Cette méthode n'impose aucune restriction sur le pas τ et il y a lieu de s'attendre à ce qu'en l'utilisant on atteindra plus rapidement l'état stationnaire limite, qui est la solution de notre problème (231), (232).

C'est pourquoi nous allons revenir de nouveau aux formules (226), (227) en ajoutant le second membre $f_{k,m}$ et nous allons les utiliser de paire avec les conditions aux limites correspondantes pour trouver la solution du problème stationnaire envisagé. Pour simplifier l'étude de la méthode, nous allons nous limiter au cas où le domaine G est un carré $0 \leqslant x, y \leqslant X$ et les pas du réseau sont égaux entre eux $h_x = h_y = h$. La transition de la ν-ième itération à la $(\nu+1)$-ième consiste à résoudre le système d'équations

$$\frac{\tilde{u}_{k,m} - u^{(\nu)}_{k,m}}{\tau/2} = \frac{\tilde{u}_{k+1,m} - 2\tilde{u}_{k,m} + \tilde{u}_{k-1,m}}{h^2} + \frac{u^{(\nu)}_{k,m+1} - 2u^{(\nu)}_{k,m} + u^{(\nu)}_{k,m-1}}{h^2} - f_{k,m} \tag{249}$$

par rapport à $\tilde{u}$, puis le système d'équations

$$\frac{u^{(\nu+1)}_{k,m} - \tilde{u}_{k,m}}{\tau/2} = \frac{\tilde{u}_{k+1,m} - 2\tilde{u}_{k,m} + \tilde{u}_{k-1,m}}{h^2} + \frac{u^{(\nu+1)}_{k,m+1} - 2u^{(\nu+1)}_{k,m} + u^{(\nu+1)}_{k,m-1}}{h^2} - f_{k,m} \tag{250}$$

par rapport à $u^{(\nu+1)}$. Les indices k, m prennent des valeurs de 0 à $K = X/h$, les grandeurs $\tilde{u}$, $u^{(\nu+1)}$ tout comme $u^{(\nu)}$ sont données sur la frontière, c'est-à-dire pour $k, m = 0$,

K et à chaque point interne correspond une paire d'équations (249), (250).

Pour ce qui est de l'algorithme de résolution des systèmes (249) et (250), nous l'avons déjà étudié, c'est la vitesse de convergence du processus itératif qui nous intéresse maintenant. Cette étude est analogue à celle qui a été faite ci-dessus dans le cas d'une itération simple. Posant $u^{(\nu)} = u + \delta^{(\nu)}$ on obtient pour l'erreur $\delta^{(\nu)}$ le même problème aux différences (249), (250) avec $f = 0$ et des conditions aux limites nulles. Les fonctions discrètes $\delta_{k,m}$ (246) avec

$$\varphi_p = p\pi/K,\ \psi_q = q\pi/K;\ p,\ q = 1,\ 2,\ \ldots,\ K-1 \quad (251)$$

seront de nouveau les fonctions propres de l'opérateur d'itérations. Les valeurs propres λ déterminant la vitesse de convergence s'expriment par la formule (228) avec $\varphi = \varphi_p$, $\psi = \psi_q$ c'est-à-dire elles sont des produits de facteurs de même type $\lambda = \lambda_p \lambda_q$. Ecrivons l'un d'eux

$$\lambda_p = \frac{1 - \frac{2\tau}{h^2} \sin^2 \frac{\varphi_p}{2}}{1 + \frac{2\tau}{h^2} \sin^2 \frac{\varphi_p}{2}}. \quad (252)$$

Le second facteur λ_q s'obtient à partir du premier en remplaçant φ_p par ψ_q et par conséquent

$$\max |\lambda| = \max |\lambda_p| \max |\lambda_q| = \max |\lambda_p|^2. \quad (253)$$

Il est évident que $\max |\lambda| < 1$ pour n'importe quelle valeur positive du paramètre τ, c'est-à-dire le processus itératif converge toujours. Mais pour pouvoir estimer la vitesse de convergence, on aura besoin d'une estimation la plus exacte possible de la grandeur $|\lambda|$.

Comme $Kh = X$, φ_p a pour limites de variation

$$h\pi/X \leqslant \varphi_p \leqslant \pi - h\pi/X, \quad (254)$$

et pour λ_p pour h petit, on a l'estimation suivante (fig. 22)

$$\frac{1-\tau\frac{\pi^2}{2X^2}[1+O(h^2)]}{1+\tau\frac{\pi^2}{2X^2}[1+O(h^2)]} \geqslant \lambda_p \geqslant \frac{1-\tau\frac{2}{h^2}[1-O(h^2)]}{1+\tau\frac{2}{h^2}[1-O(h^2)]}. \quad (255)$$

Nous avons besoin que max $|\lambda_p|$ soit minimal. Nous avons à notre disposition le paramètre τ, qui représente le temps fictif; il semble que plus grand est τ, meilleure doit être la convergence, car on atteint alors plus rapidement l'état limite. En effet, lorsque τ augmente, le premier membre de (255) décroît. Cependant, le second membre de (255) tendra alors vers -1, ce qui aura pour effet de ralentir la convergence. Cet effet peut s'expliquer par ce que, bien que pour τ grand on se déplace plus rapidement dans le temps, ce déplacement est plus grossier, l'imprécision de la solution obtenue surpasse sa variation.

Il est évident que la valeur optimale de τ sera celle pour laquelle les premier et second membres de (225) sont égaux en module. En négligeant les infiniment petits d'ordre supérieur à $O(h^2)$, on obtient l'équation suivante pour τ:

$$\frac{1-\tau\frac{\pi^2}{2X}}{1+\tau\frac{\pi^2}{2X^2}} = -\frac{1-\tau\frac{2}{h^2}}{1+\tau\frac{2}{h^2}},$$

dont la solution est

$$\tau = hX/\pi. \quad (256)$$

En substituant cette valeur de τ dans l'inégalité (255), on obtient l'estimation

$$\max|\lambda_p| = 1 - h\pi/X + O(h^2),$$

c'est-à-dire, en vertu de (253),

$$\max|\lambda| = 1 - 2h\pi/X + O(h^2),$$

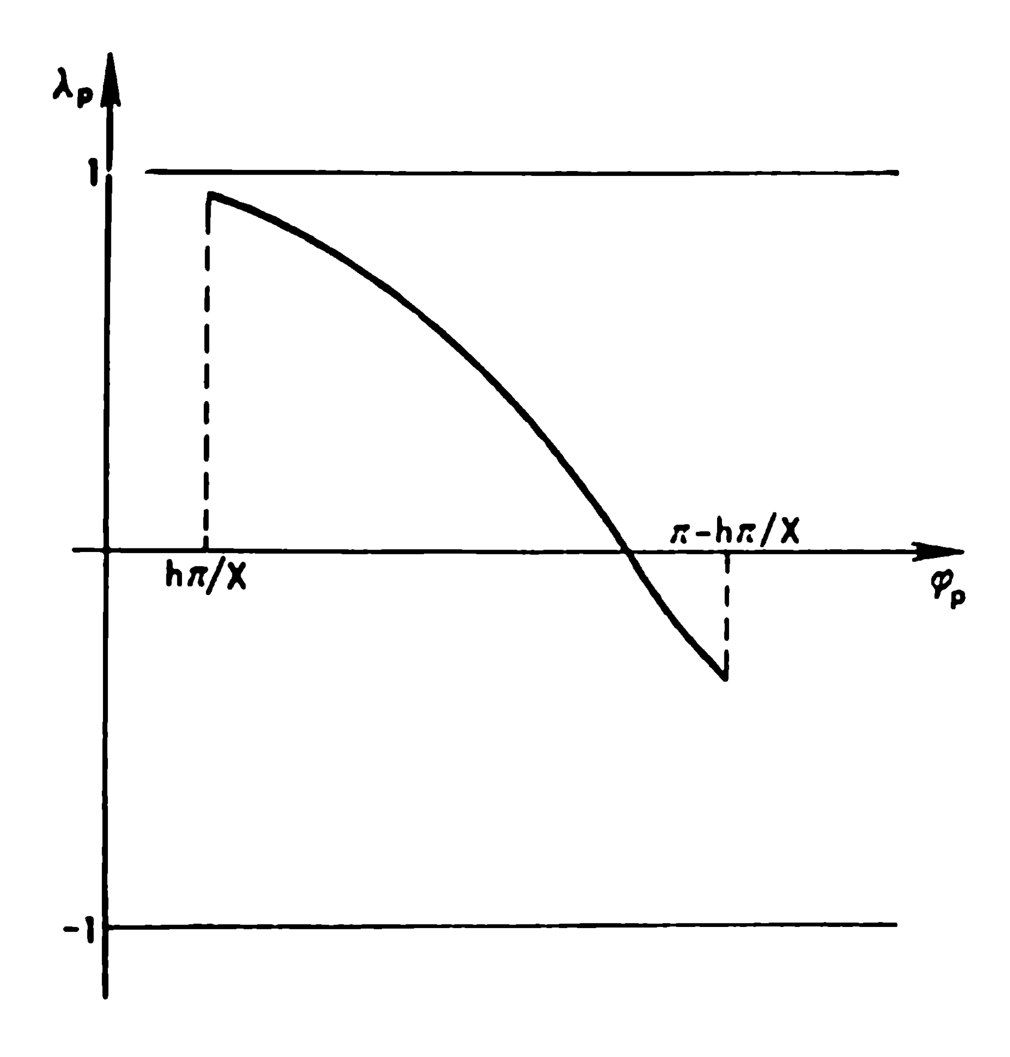


Fig. 22

et la grandeur $\varkappa$ caractérisant la vitesse de convergence se trouve être de l'ordre de h

$$\varkappa \sim h\frac{2\pi}{X}. \tag{257}$$

Ainsi le processus itératif utilisant la méthode des directions alternées avec $\tau \sim h$ a une vitesse de convergence dix fois meilleure que la simple itération où $\varkappa \sim h^2$ (248).

Si l'on développe l'erreur δ en composantes qui sont des harmoniques de la fonction (246), (251), $\lambda_p\lambda_q$ sera alors le coefficient d'amortissement de la composante pour une itération. La formule (252) et la fig. 22 montrent que les différentes composantes sont amorties différemment. Les

plus amortis sont les harmoniques de fréquences φ_p pour lesquels

$$\frac{2\tau}{h^2}\sin^2\frac{\varphi_p}{2}\sim 1.$$

Il est évident que le choix de τ permet de régler la gamme des fréquences. Ainsi pour la valeur mentionnée de τ (256) ce sont les fréquences moyennes. Mais leur amortissement trop important n'est pas intéressant car la vitesse de convergence est déterminée par les coefficients d'amortissement des fréquences extrêmes φ_p, $\psi_q \sim h$ et φ_p, $\psi_q \sim \pi$. Ceci porte à penser qu'il serait bon d'utiliser dans les différentes itérations différentes valeurs de τ, ceci afin que d'assurer un amortissement uniforme de toutes les fréquences. Ainsi, en utilisant des suites $\tau = \tau^{(\nu)}$ spécialement élaborées, on arrive à obtenir des méthodes dont la vitesse de convergence est encore plus grande que (257), par exemple (voir problème 2)

$$\varkappa \sim \frac{1}{\ln(1/h)}. \tag{258}$$

Nous n'allons pas nous arrêter sur ces méthodes.

Toutes les méthodes itératives de résolution des systèmes d'équations aux différences (231) ont une vitesse de convergence faible qui décroît plus ou moins rapidement lorsque h augmente: $\varkappa \to 0$ pour $h \to 0$. Néanmoins elles sont plus avantageuses que les méthodes directes non itératives. Nous allons donner une estimation du nombre d'opérations arithmétiques N nécessaires à la résolution du problème.

Commençons par la méthode générale d'exclusion. Lorsqu'on l'utilise, le nombre d'opérations nécessaires est de l'ordre du cube du nombre d'inconnues. Ce dernier étant de l'ordre de $1/h^2$, on a

$$N \sim 1/h^6. \tag{259}$$

Dans le paragraphe précédent nous avons décrit la méthode du balayage matricielle qui est, de toute évidence,

applicable dans le cas présent. Pour cette méthode il faut

$$N \sim 1/h^4 \tag{260}$$

opérations arithmétiques car pour inverser une matrice d'ordre $1/h$ il faut $1/h^3$ opérations, et on a en tout $\sim 1/h$ de telles matrices.

Les méthodes itératives donnent la solution approchée du système, c'est pourquoi il y a lieu d'estimer le nombre d'opérations nécessaires pour diminuer l'erreur d'un nombre donné de fois. Lors d'une itération l'erreur décroît de $1 - \varkappa$ fois. Si l'erreur de l'approximation initiale est prise égale à l'unité, après n itérations elle sera égale à

$$(1-\varkappa)^n = \varepsilon.$$

Par conséquent pour que la précision soit égale à ε, il faut

$$n = \frac{\ln \varepsilon}{\ln(1-\varkappa)} \sim \frac{\ln(1/\varepsilon)}{\varkappa}$$

itérations. Dans les méthodes envisagées le nombre d'opérations arithmétiques pour une itération est de l'ordre du nombre de points du réseau, c'est-à-dire $\sim 1/h^2$. Par conséquent, le nombre total d'opérations est

$$N \sim \frac{\ln(1/\varepsilon)}{\varkappa h^2}.$$

En y substituant les différentes valeurs de $\varkappa$ à partir de (248), (257), (258), on obtient : pour une itération simple

$$N \sim \ln(1/\varepsilon)/h^4, \tag{261}$$

pour la méthode des directions alternées

$$N \sim \ln(1/\varepsilon)\, h^3 \tag{262}$$

et dans le cas d'un choix spécial de τ

$$N \sim \ln(1/\varepsilon) \ln(1/h)/h^2. \tag{263}$$

En comparant (259), (260) avec (261) à (263), on voit l'avantage des premières. Il est évident que l'économie des méthodes itératives provient de ce que lorsqu'on les utilise on arrive à tenir compte au maximum des particularités du système d'équations.

Problèmes

1. Déterminer la valeur optimale du paramètre τ lors du processus itératif suivant les formules du type (249), (250) dans le cas où le domaine G est un rectangle de côtés X, Y et les pas h_x, h_y ne sont pas égaux entre eux.

2. Montrer qu'il est possible d'obtenir la vitesse de convergence des itérations donnée par la formule (258). A cet effet étudier le processus itératif (249), (250) à τ variable:

$$\tau^{(\nu)}=\tau^{(1)}z^{\nu-1}, \qquad \nu=1,\ 2,\ \ldots,\ n,$$
$$\tau^{(\nu)}=\tau^{(\nu-n)}, \qquad \nu=n+1,\ n+2,\ \ldots$$

Choisir les valeurs $\tau^{(1)}$ et n de telle sorte que la grandeur

$$\xi_p^{(\nu)}=\frac{2\tau^{(\nu)}}{h^2}\sin^2\frac{\varphi_p}{2},$$

figurant dans l'expression de λ_p (252) pour tout p admissible se trouve pour un certain $\nu=\nu(p)$ dans l'intervalle entre $\sqrt{z}$ et $1/\sqrt{z}$. La valeur $\lambda_p^{(\nu)}$ correspondant à ce ν doit vérifier l'inégalité

$$|\lambda_p^{(\nu)}|<\left|\frac{1-\sqrt{z}}{1+\sqrt{z}}\right|,$$

qui peut être utilisée pour estimer la valeur moyenne λ_p pour un cycle de n itérations et par conséquent la vitesse de convergence.

3. Etudier la possibilité d'application des méthodes mentionnées dans le problème 5 du § 11 au cas des problèmes stationnaires.

4. Proposer un algorithme de calcul pour la solution de l'équation (229) dans les domaines de géométrie compliquée (figure formée de rectangles, cercle et autres) pour différentes conditions aux limites.

5. Elaborer et étudier les schémas aux différences pour la résolution des systèmes d'équations

$$\frac{\partial U}{\partial x}+\frac{\partial V}{\partial y}=f(x,\ y),\ \frac{\partial V}{\partial x}-\frac{\partial U}{\partial y}=g(x,\ y)$$

dans le rectangle $0 \leqslant x \leqslant X$, $0 \leqslant y \leqslant Y$ avec les conditions aux limites sur les côtés

$$U(0,\ y)=\alpha(y), \qquad U(x,\ Y)=\beta(x),$$
$$V(x,\ 0)=\gamma(x), \qquad V(X,\ y)=\delta(y).$$

En particulier, envisager l'algorithme itératif utilisant des équations aux différences du type

$$\tilde{u}_{k,\ m-1/2}+\tau\frac{\tilde{u}_{k,\ m-1/2}-\tilde{u}_{k-1,\ m-1/2}}{h_x}=$$
$$=u^n_{k,\ m-1/2}-\tau\frac{v^n_{k-1/2,\ m}-v^n_{k-1/2,\ m-1}}{h_y}+\tau f_{k-1/2,\ m-1/2},$$

$$\tilde{v}_{k-1/2,\ m}-\tau\frac{\tilde{v}_{k+1/2,\ m}-\tilde{v}_{k-1/2,\ m}}{h_x}=$$
$$=v^n_{k-1/2,\ m}-\tau\frac{u^n_{k,\ m+1/2}-u^n_{k,\ m-1/2}}{h_y}-\tau g_{k,\ m},$$

$$u^{n+1}_{k,\ m-1/2}+\tau\frac{v^{n+1}_{k-1/2,\ m}-v^{n+1}_{k-1/2,\ m-1}}{h_y}=$$
$$=\tilde{u}_{k,\ m-1/2}-\tau\frac{\tilde{u}_{k,\ m-1/2}-\tilde{u}_{k-1,\ m-1/2}}{h_x}+\tau f_{k-1/2,\ m-1/2},$$

$$v^{n+1}_{k-1/2,\ m}+\tau\frac{u^{n+1}_{k,\ m+1/2}-u^{n+1}_{k,\ m-1/2}}{h_y}=$$
$$=\tilde{v}_{k-1/2,\ m}+\tau\frac{\tilde{v}_{k+1/2,m}-\tilde{v}_{k-1/2,\ m}}{h_x}-\tau g_{k,\ m},$$

$$k=1,\ 2,\ \ldots,\ K;\quad m=1,\ 2,\ \ldots,\ M;$$

$$h_x=\frac{X}{K+{}^1/_2},\qquad h_y=\frac{Y}{M+{}^1/_2},$$

où n est le numéro de l'itération, $\sim$ désigne les valeurs intermédiaires des inconnues, τ un certain paramètre.

Table des matières

À NOS LECTEURS

Les Editions Mir vous seraient très reconnaissantes de bien vouloir leur communiquer votre opinion sur ce livre, sa traduction et sa présentation, ainsi que toute autre suggestion.

Notre adresse:

Editions Mir 2, Pervi Rijski péréoulok, Moscou, I-110, GSP, U.R.S.S.



